2013

Does Attainment of Piaget's Formal Operational Level of Cognitive Development Predict Student Understanding of Scientific Models?

Richard Dennis Lahti
The University of Montana

DOES ATTAINMENT OF PIAGET'S FORMAL OPERATIONAL LEVEL OF COGNITIVE DEVELOPMENT PREDICT STUDENT UNDERSTANDING OF SCIENTIFIC MODELS?

By

RICHARD DENNIS LAHTI II

M.S. Science Education, Montana State University, Bozeman, MT, August 10, 2001
B.S. Chemistry, Michigan State University, East Lansing, MI, August 19, 1994

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctorate of Education
in Curriculum and Instruction

The University of Montana
Missoula, MT

December, 2012

Approved by:

Sandy Ross, Associate Dean of The Graduate School
Graduate School

Darrell Stolle, Co-Chair
Department of Curriculum and Instruction

Mark Cracolice, Co-Chair
Department of Chemistry and Biochemistry

David Erickson
Department of Curriculum and Instruction

Lisa Blank
Department of Curriculum and Instruction

Georgia Cobbs
Department of Curriculum and Instruction

Lahti, Richard, Ed.D., Fall 2012, Curriculum and Instruction

DOES ATTAINMENT OF PIAGET'S FORMAL OPERATION LEVEL OF COGNITIVE DEVELOPMENT PREDICT STUDENT UNDERSTANDING OF SCIENTIFIC MODELS?

Co-Chairperson: Darrell Stolle

Co-Chairperson: Mark Cracolice

Knowledge of scientific models and their uses is a concept that has become a key benchmark in many of the science standards of the past 30 years, including the proposed Next Generation Science Standards. Knowledge of models is linked to other important nature of science concepts such as theory change which are also rising in prominence in newer standards. Effective methods of instruction will need to be developed to enable students to achieve these standards. The literature reveals an inconsistent history of success with modeling education. These same studies point to a possible cognitive development component which might explain why some students succeeded and others failed. An environmental science course, rich in modeling experiences, was used to test both the extent to which knowledge of models and modeling could be improved over the course of one semester, and more importantly, to identify if cognitive ability was related to this improvement. In addition, nature of science knowledge, particularly related to theories and theory change, was also examined. Pretest and posttest results on modeling (SUMS) and nature of science (SUSSI), as well as data from the modeling activities themselves, was collected. Cognitive ability was measured (CTSR) as a covariate. Students' gain in six of seven categories of modeling knowledge was at least medium (Cohen's d >.5) and moderately correlated to CTSR for two of seven categories. Nature of science gains were smaller, although more strongly correlated with CTSR. Student success at creating a model was related to CTSR, significantly in three of five sub-categories. These results suggest that explicit, reflective experience with models can increase student knowledge of models and modeling (although higher cognitive ability students may have more success), but successfully creating models may depend more heavily on cognitive ability. This finding in particular has implications in the grade placement of modeling standards and curriculum chosen to help these students, particularly those with low cognitive ability, to meet the standards.

DEDICATION

This dissertation is dedicated to my wife, Katharine and my children Erika and Sonja. May the destination justify the journey, and may the lost time together be paid back with interest after graduation.

ACKNOWLEDGEMENTS

There are a number of people and institutions I wish to acknowledge and thank for the support that they have given me during this process. First, I thank Dr. Darrell Stoll for not giving up on me even when I was thinking of giving up on the study myself. Second, I thank Dr. Mark Cracolice. The lessons taught through constructivism take longer, but sink in more because of it. I would like to thank the rest of my dissertation committee, Dr. David Erickson, Dr. Lisa Blank, Dr. Georgia Cobbs, and Dr. Fletcher Brown, for their time and patience in this process, from agreeing to sit on a committee for a student that that some of you had never met to filling in last minute as a sabbatical replacement, so that I could complete my degree this semester. In particular I thank Dr. Erickson for his careful edits on the proposal and dissertation. I would also like to acknowledge the Center for Learning in the West and the National Science Foundation (grant # M27875) for their generous support, and the following people at CLTW who made this opportunity possible: Dr. Cobbs, Dr. Elizabeth Swanson, and Dr. Libby Krussel-Knot. Thank you very much to Dr. Curt Doetkott and especially Dr. Wendy Troop-Gordon, both of North Dakota State University, for statistical support without which this study would have never been completed. Finally, I would like to thank the administration at MSUM for their support and patience in a process that took longer than it should but is hopefully worth it in the end.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER ONE

THE PROBLEM TO BE STUDIED

Background

The reform movement of the 1980s.

For the non-scientists, science can be an intimidating, exclusionary field. Turner (2008) reports in his history of scientific literacy that the 1980s were to spawn several movements aimed at increasing science awareness for the public at large in Canada, the United States, and the United Kingdom. Terms such as science literacy and science, technology, and society came to the forefront of science education discussions during this time period. Techniques such as authentic scientific inquiry replaced traditional science education. *A Nation at Risk* (National Commission on Excellence in Education, 1983) spurred curricular reforms aimed at educating all society about science, instead of just the academic elite. Many of these changes came about because of global economic pressures and increasingly scientific political issues such as energy and genetic modification (Turner, 2008). Several new assessments for measuring scientific understanding were created as a result. And, as scientific models are an integral part of science, it is not surprising that the seminal work in understanding scientific models by Grosslight, Unger, Jay, and Smith (1991) followed shortly thereafter.

One such science reform document from the 1980s was *Science for all Americans.* *Science for all Americans* (AAAS, 1989) sets forth guidelines describing what every scientifically literate American should know about science before Haley's Comet returns in the year 2061. In addition to specifying what knowledge Americans need to have, it also establishes

why this knowledge is important and important issues surrounding this knowledge. Knowledge about models and modeling is a topic worthy of three pages in *Science for all Americans*, Chapter 11: Common Themes (AAAS, 1989). These pages provide an overview of models, their types, and their relationship to scientific theories. Included in the elaboration are many of the difficulties learners of all ages have experienced regarding models and theories.

Nature of models.

While the definitions of the vocabulary regarding models will be explored in greater depth in the definition section and review of the literature, a few basic concepts must be introduced at this point. A model is, by definition, a representation of a target (concept, object, phenomenon, relationship, system); thus, the first important consideration is that the model is not the target. Therefore, since it is not the target, it must differ from the target in at least one way (Grosslight, Unger, Jay, & Smith, 1991). A second important consideration is that a model is a representation, and no model is ever *correct* (Harrison & Treagust, 2000). One particular model may provide greater accuracy than another model, may provide similar accuracy with less complexity, or may simply be more suited to a particular situation, but a model cannot give a completely accurate representation of a target or a phenomenon under all conditions (Harrison & Treagust, 2000).

Grosslight et al., (1991) propose three levels of modeling understanding that will be used throughout this work. At the first level, typical of many students, models are used *to show*. The primary focus is on attaining the most exact physical representation of the phenomenon (in this case, typically an object). At the second level, typical of many teachers, models are used *to communicate*. There is less emphasis on achieving an exact replica because it is understood that simplifications of unimportant aspects and emphasis of the important aspects of the phenomenon

in the model can result in better communication. Also at this level, the emphasis moves from the physical representation (form), towards capturing the behavior of the phenomenon (function). At the third, or expert, level the function of models is *to predict*. Models allow scientists to generate testable hypothesis. These levels are not mutually exclusive; an expert can still use models in the other two ways as appropriate.

A scientific model differs from the 14 lay definitions (Gove, 1981) of models in that it contains unseen and postulated components (Van Driel & Verloop, 1999). The lay definitions include "drawings to scale," "thing that exactly resembles another," "usually miniature three-dimensional representation," "pattern," "a person or thing regarded as worthy of imitation," "archetype," "one who is employed to display clothes," and "a specific type or design" (Gove, 1981, p. 1451). The 14th and final lay definition is closer to a scientific model in that it talks about "relationships" between parts and that a model helps to "visualize often in a simplified way something that cannot be directly observed" (Gove, 1981, p. 1451), which seems consistent with a level two conception of model. Fully scientific models (level three) are also called hypothetical deductive models. For this study a model will be defined as a representation (physical, conceptual, or mathematical) of a target phenomenon, intended to communicate significant aspects of the phenomenon and to form hypotheses about the phenomenon.

How do student conceptions of models form?

Student conceptions of models come from everyday life as well as science class. While much of what students learn about models from other sources is applicable to scientific models, it is the aspects of models that do not pertain to scientific models that lead to the greatest difficulty.

There are several reasons for student confusion about models, although much of the difficulty may be attributed to students' first interactions with models, typically models not of the scientific variety. In a child's life, model airplanes and other toys are meant to be concrete representations of the object they are imitating, and are rarely used for the higher purposes that scientific models may be used for, such as improving communication about the phenomenon, and rarer still for testing hypotheses. Grosslight et al. (1991) found that most seventh and eleventh grade students display a naïve realist epistemology and see models as concrete representations of reality.

Students rarely model in the scientific sense, but when students create models, they typically start with form over function (Lehrer & Schauble, 2003). This tendency would seem to be related to their disproportionate exposure to physical models, and also any informal modeling they may have done at play (when children play war, for example, the sticks that have handles and look like guns become guns, those that look like swords become swords, and so forth, although none are functionally correct). Therefore, their initial models start by looking like what they are modeling and proceed on to functionality (Lehrer & Schauble, 2003).

Since the behavior and predictive power of a scientific model is more important than its form, it is natural that students have difficulty with the nature of scientific models. A lack of modeling in the curricula contributes to this problem. Although Lehrer and Schauble (2003) started with an emphasis on the form of the model, after instruction in models and modeling, they found that students began to appreciate the other uses of a model, first in using the model to aid communication and finally in using the model as a way to represent and test ideas. Unfortunately, science curricula to this point have not emphasized modeling (Justi & Gilbert, 2002a; Van Driel & Verloop, 1999), and teachers often teach models poorly, if at all (Justi &

Gilbert, 2002b). Saari and Viiri (2003) add that if the gains in modeling knowledge are to be more or less permanent, modeling needs to continue to be part of the curriculum not a single experience, or the students will regress back to their previous modeling level. Thus, even if there exists pockets of good modeling instruction, the lack of consistency will hamper the effort. The current proposed draft of the Next Generation Science Standards (2012), however, have models as one of seven crosscutting concepts across all science content. These new standards are based on *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* the Committee on Conceptual Framework for the New K-12 Science Education Standards of the National Research Council (2012). This consistent, integrated approach could provide the consistency that was lacking in older curricula.

Salient Student Variables

Cognitive development.

There is mixed evidence to suggest that Piagetian developmental level would be related to the ability to model. In her critique of Piaget, Driver (1978) claims his very definition of formal thought is "the existence of 'integrated structures' that could be modeled mathematically" (p. 55). However, interpretation of this statement is heavily dependent upon the definition of model and modeling. Since the variable cognitive development is confounded with the variable age, it is helpful to first examine age as a variable related to modeling. The research studies, discussed in more detail in the review of the literature, to teaching modeling had success, but those that recorded individual student achievement found failure as well.

Age and/or cognitive development as variables.

As the literature review will reveal, age and/or cognitive development appear to be significant variables relating to the relative success and failure of various modeling studies and

curricula. Because chronological age and cognitive development are related but not identical, the varying levels of success observed in the studies could correspond to differences in cognitive development among the students in the sample.

Why is Modeling Important?

Science is intimately connected with models and modeling

Students need to understand models in order "to learn science … to learn about science … and to learn how to do science" (Justi & Gilbert, 2002a, p. 370). This partial quote sums up the central role that many feel scientific models play in science, and specifically, in science education. Yet, there is at least as much confusion among students regarding scientific models as there is regarding other aspects of the nature of science.

Models resemble theories in many ways. For instance, models and theories are both imperfect reflections of reality. Because an understanding of the nature of a scientific theory is of importance to being scientifically literate (AAAS, 1989), utilizing similarities between models and theories would seem to be an important approach to understanding both. This similarity between model and theory makes sense since the words "theory" and "model" are used interchangeably in some science writing. For example, in *The Making of the Standard Model*, Hooft (2007) states, "The standard model of particle physics is more than a model. It is a detailed theory that encompasses nearly all that is known about the subatomic particles and forces in a concise set of principles and equations" (p. 271). The first two sentences could be simplified to "The … model … is a … theory."

What, then, is a scientific theory? Of the five unique uses of the word "theory" in the English language, it is unfortunate that the lay use synonymous with "conjecture, speculation, and supposition" (Gove, 1981, p. 2371) is so at odds with the scientific use. In science, far from

mere speculation, a theory is a powerful set of ideas that are used to explain a variety of observations, to relate previously unrelated phenomenon, and to provide accurate predictions (AAAS, 1989), or "powerful tools" that "have the potential to lead to new knowledge" (NSTA, 2003). These definitions both point to the predictive nature of theories, yet neither provides for another essential characteristic of a scientific theory. Initially, a scientific theory presupposes an invisible or unseen mechanism. For instance, Mendel invented unobservable genotypes to explain the observed phenotypes in his theory of genetic inheritance (Lawson, Alkoury, Benford et al., 2000). Lawson et al. define "theoretical concepts" (p. 997) as "only indirectly testable" and "function as explanations for events that need causes, but for which no causal agent can be perceived," and points out that these theoretical concepts included many now familiar constructs such as "photons, electrons, atoms, molecules and genes."

Models, hypotheses, laws, and theories.

Since models are related to theories, they are also related to hypotheses and laws, two other important concepts in the nature of science. Students have several misconceptions involving hypotheses. One misconception discussed in Windschitl and Thompson (2006) is the oversimplified idea of a hypothesis. While most students can parrot the definition *educated guess* for a hypothesis, students rarely connect *educated* with grounding in a theory or model. Students also have difficulty with the process of how a scientific hypothesis becomes a theory or law, and how these theories can be refined over time. Worse yet, many students subscribe to a hierarchical view that hypotheses become theories which then become laws. These misconceptions are barriers to understanding an important part of the nature of science.

In reality, models are useful in enabling scientists to test hypotheses and theories (Treagust, Chittleborough, & Mamiala, 2002). A successful experiment using a model helps to

support both the model and the theory from which the model was constructed. An experiment that yields results contrary to the hypothesis can show that either the model, the theory underlying the model, or both is in need of revision.

Closing the gap.

With such a large gap existing between students' initial conceptions about models and science and full scientific conception, it may be ambitious to expect students to move completely to an expert conception in a short period of time such as a one-semester class. For the non-scientists that make up the majority of society (and the sample in this study), understanding the unseen aspect of both theories and models may not be the primary goal, as it is not discussed in Science for All Americans (AAAS, 1989), nor in the National Science Teachers Association position statement on the Theory of Evolution (NSTA, 2003). While students in this study may not propose new models and theories containing "theoretical concepts" and therefore may not reach a full scientific understanding of models and theories, this study will attempt to improve student conceptions of both by focusing on the aspects of theory building and modeling that are more directly accessible. These aspects include deriving of testable hypotheses, purposive selecting of components to include in a model, compromising between complexity and accuracy, and developing models through iterations (Van Driel & Verloop, 1999). This approach is consistent with the approach in Clement (2000) of teaching a target model through a set of increasingly authentic intermediary models and the model of modeling framework presented in Justi and Gilbert (2002a). In short, it may represent a step in the right direction.

Furthermore, revising a mathematical model, such as a spreadsheet, is a fairly simple process that students can experience directly. Since model refinement is similar to theory

building, it is possible that experience refining models could give increased understanding of theory building.

Problem to be Studied

To this point, three important factors have been discussed; developmental level, the nature of science, and scientific models. Although there are strong connections between (a) the nature of science and scientific models and some connections between (b) models and cognitive development and (c) the nature of science and cognitive development, the review of the literature will reveal no study that examines the relationship between all three variables. Specifically, cognitive development seems to be a limiting factor in students' ability to understand models, especially those involving unseen agents. Theories are built on just such models. Therefore, the extent to which cognitive development (in the Piagetian sense) predicts improved understanding of the nature of science through a modeling curriculum is not clearly understood.

Theoretical Orientation

Several theoretical perspectives will influence this study. First, one of the central tenets of modeling education is that by creating a useful model, students construct a deep understanding of all aspects of the system to be studied. The teacher provides the background information, but it is up to the student modelers themselves to identify the relevant information, to quantify it, to create the model, and finally to go through the process of verification of their model against new data (Penner, Lehrer, & Schauble, 1998). If students perform this activity in a setting where exchange of information is allowed, constructivism becomes an important theoretical perspective. As different students may arrive with different data sets and backgrounds, they may perceive the strengths and weaknesses of particular models differently. The sharing of these experiences and performance of each model on new data sets helps students decide for

themselves how and why their models should be modified in light of the results (Penner et al., 1998). Thus, this approach is aligned with the constructivist perspective and far different from the traditional teaching paradigm of concept presentation and subsequent data collection and verification.

Some modeling literature specifically addresses the Piagetian level of students and modeling. As discussed previously, true hypothetical scientific models with unseen causative agents would seem to support a post-formal level of development (Lawson et al., 2007). Even if students were provided concrete causative agents, reasoning abstractly about them would require formal operational level (Lawson, Banks, & Logvin, 2007). There are significant parallels between the steps in model development and Piaget's stages (Lesh & Doerr, 2003; Lesh & Carmona, 2003). At the final stage, students in the modeling activity should be involved in formal operational thinking, by setting up a proportion between two related multiplicative relationships (Lesh & Doerr, 2003). Another example of overlap is Piaget's notion of accommodation (changing mental structures in light of contradictory evidence) which is similar to model revision when the model fails to adequately explain some portion of the data. Although some aspects of Piaget's theories have fallen under criticism (Driscoll, 1994; Driver, 1978), they provide the most common and useful framework to begin discussions on cognitive development. Thus, this research is concerned with some of the issues central to the debates about Piaget's work.

Purpose

The purpose of this study is to determine if a modeling curriculum entailing the repeated utilization and comparison of multiple mathematical models in an environmental science class will have an effect on student understanding of modeling and the nature of science. More

importantly, does cognitive development, as described by Piagetian theory, predict a student's ability to benefit from said curriculum?

Significance

Although the literature review will reveal that student conceptions of models and modeling have been studied in some detail, there appears to be a new emphasis on modeling in science. The university of the author has a new liberal studies program that directly lists modeling as a goal (MSUM, 2006). At the K-12 level, the Minnesota Academic Standards in Science assessed by the Science Minnesota Comprehensive Assessment II (MCA II) (MDOE, 2005; MDOE 2006) and dictated by No Child Left Behind (NCLB) brings modeling instruction to the forefront. Thus, because of these mandates, the much researched questions of *can* or *should* modeling be used to improve science instruction becomes the different question of *how to effectively teach modeling?* Given the dearth of modeling in the traditional K-12 science curricular materials (Justi & Gilbert, 2002a), and in pre-service teacher education (Cullin & Crawford, 2003), it becomes important to determine if a particular approach is, first, successful at teaching the concepts and second, if the success and failure of such an approach is related to students' cognitive developmental level.

Research Questions

Research question.

Is attainment of the formal operational Piagetian level of cognitive development necessary for a model-based environmental science curriculum to increase students' understanding of models and the nature of science?

Sub-questions.

1. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations), and is that improvement related to Piagetian level?
2. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the utility of models (communication, simplification for study, prediction), and is that improvement related to Piagetian level?
3. Does a curriculum emphasizing student comparison, refinement, and creation of models improve student understanding of the relationship between models, theories, and the scientific method (models operationalize theories, allowing them to be tested with the scientific method) , and is that improvement related to Piagetian level?

Hypotheses

Null hypotheses.

There will be no significant difference at the p =.05 level in student understanding of models nor understanding of the nature of science before and after completing a semester of the model-laden environmental science curriculum. Moreover, there will be no significant difference at the p =.05 level between any normalized gain between the pretest and posttest in modeling and/or nature of science knowledge between students with differing cognitive development as measured by the Classroom Test of Scientific Reasoning (CTSR). (This curriculum included exposure to authentic model use, critique and modification of existing models, comparison of multiple models of the same system, analysis of the conscious choices

that shape models, and finally construction of their own models and using these models to answer questions.)

Alternative hypotheses.

There will be statistically significant difference between students' modeling knowledge and/or nature of science scores on the posttest as compared to the pretest. This difference will also show a normalized gain of greater than 0.5 (medium effect). When any gains in modeling and/or nature of science knowledge are correlated to the CTSR score, students with larger CTSR scores, and thus more developed formal reasoning, will have statistically greater gains than students with lower levels of development. There will be a correlation of at least 0.5 between formal knowledge score and gains from the pretest to posttest.

Variables

Independent variable.

Scores obtained by students on the Classroom Test of Scientific Reasoning was the independent variable. This pretest score was interval level data that was analyzed as such.

Dependent variables.

A modified version of the Student Understanding of Science and Scientific Inquiry Questionnaire (SUSSI) which contained both Likert-scale and free-response questions, was used as both a pretest and a posttest. Minor modifications were made by the author to eliminate ambiguity and/or increase alignment with university language (see Appendix A for modifications). The free response was scored with a rubric (see Appendix B for a copy of the rubric) tested on peers of the students of this study, and scored at least twice, and at least once by someone other than the author. The scores on this are interval level data, and were analyzed as such. This score was the primary quantitative measure of nature of science knowledge.

.

A modified version of the Student Understanding of Models in Science (SUMS) instrument was the primary quantitative measure of modeling knowledge (see Appendix A for a copy of the test). This test was exclusively Likert-scale in its original formal. The modifications included language changes to reduce ambiguity, reversal of some answers (so that strongly agree was not always the best answer), and the addition of free-response questions. This variable is interval level data.

Qualitative measures.

In addition to these quantitative sources, qualitative sources of information were used to attempt to explain the numerical results. Some of these sources were more structured and intentional, while others looked at emerging trends in the data. A more complete handling of these measures is detailed in Chapter Three.

Two sets of student interviews were conducted. A subset of students was interviewed following the pretests and posttests in an attempt to determine if the written instruments (SUSSI and SUMS) accurately gauged student knowledge. Furthermore, this interview process was repeated after the posttest to determine the reasons that answers were changed from pretest to posttest. The prompts used are listed in Appendix C: Interview Protocols.

The individual student answers on modeling assignments constituted another important source of information. The individual questions were both scored directly on a scale of one to three and analyzed holistically for emerging trends as per Creswell (2003).

The final modeling projects were also to be scored like the reflections of the small modeling assignment, above, first on the level of modeling (one through three) and second on emerging trends. In reality, the models were broken down into five sub-scores to reflect

different aspects of the modeling process. These sub-scores were model selection, model integration, checking the model against data, using the model to test/create a hypothesis, and finally, the overall level of the model.

Definitions

The following definitions are used in this paper:

Developmental Levels: Two of Piaget's cognitive developmental levels, and one further level, are of interest in this study.

Concrete operations. The student is able to conserve and reason spatially, as well as do arithmetic with numbers that do not specifically represent concrete examples. However, the student still has difficulty with abstractions. While this stage may appear in students as young as seven, it tends to appear in pre-adolescence (Piaget & Inhelder, 1955). This may be the terminal stage of development for some people.

Formal operations. Individuals who reach this stage (not all do) are able to reason formally, including performing such tasks as compensation (i.e., if area is held constant, and length increases, width must decrease), isolation of variables, and systematically formulating and testing hypotheses (Piaget & Inhelder, 1955). While Piaget himself observed this stage beginning as early as the onset of adolescence (age 12), evidence suggests not all students reach this stage. For example Lawson, Alkoury, Benford, Clark, and Falconer (2000) found approximately half of the college students in their study had not reached the full formal level of development.

Inquiry. Inquiry is defined as an approach to teaching and learning where students learn by first interacting with data in order to develop concepts. This approach is exemplified by the 5E model of Trowbridge, Bybee, and Powell (2000).

Model. For this study a model is defined as a representation (physical, conceptual, or mathematical) of a target phenomenon, intended to communicate significant aspects of the phenomenon and to form hypotheses about the phenomenon.

Mathematical model. A mathematical model is a model as described above, where the phenomenon typically represents a system and its component parts are defined by variables and are quantifiable. Mathematical models were the primary models being analyzed and constructed. Additionally, the mathematical models were expected to show the relationship between variables in the system accurately. Within the constraints of a mathematical model, hypotheses take the form of changes to the output when certain changes to the system were made.

Scientific model. A hypothetical-deductive model with unseen causative agents as described by Lawson et al. (2007).

Delimitations

The implementation of the lessons occurred during Summer and Fall Semesters of the 2008 school year. Lessons were presented to four intact liberal-studies environmental science classes at a public, four-year university in the upper Midwest where the researcher was employed at the time.

One class in Learning Area 10: People and the Environment was required for all graduates at the institution. The learning goal for Area 10 is "To develop students' understanding of the concept of sustainability and the challenges we face in responding to environmental variables and resolving environmental problems. Students will examine how societies and the natural environment are intimately related. A thorough understanding of ecosystems and the ways in which different groups interact with their environments is the foundation of an environmentally literate individual" (MSUM, 2006).

This class is to be taken after at least one class each in mathematics, critical and multicultural thinking, natural sciences, and written and oral communications have been completed.

The results of this study should be generalizable to settings beyond other liberal studies science classes at similar colleges and universities. Other classes in science (both for majors and for non-majors) at the college as well as high school levels might find these results applicable. Furthermore, other classes where modeling is routinely done, such as computer science or mathematics, at these same levels could benefit from the study. The students in this study are not highly selected; of the new entering freshman class of 2006 (the last class for which data was available) 39.3% graduated in the bottom half of their high school class (Gill, 2007). Thus these students could represent not only a cross section of non-science majors at college, but also a good cross section of late high school abilities as well. Lessons learned from this study would be applicable to a variety of settings where a more authentic, inquiry-based method of teaching the nature of science was sought.

One should note that although this science class is at a junior level, there are no specific science prerequisites. Students entering this class must have received credit for a minimum of one science class in any discipline prior to enrolling (although this class could be taken at this institution, a previous institution of higher learning, or through advanced placement or other college credit earned in high school). The 300-level designation is not a reflection of the level of science that is experienced in the class, but rather the level of synthesis across multiple disciplines, such as mathematics (where modeling is a listed competency standard) or critical thinking (MSUM, 2006).

The lack of a large number of students of color (the campus is over 90% white) (Gill, 2007), might cause readers to question the effectiveness of this method in classrooms that are significantly different in student composition, however, there are no reasons why the results of this study would not be applicable in other settings as well, as it is built from studies from around the world and across various cultural and ability groups. Two of the more striking examples are the success with modeling approaches in mathematics for 7th grade urban, African-American students found by Lesh and Doerr (2003), and the success building spreadsheet models with algebra-resistant 15-year olds reported by Sutherland and Rojano (1993). Both studies show that a model-based approach can be successful with groups of students not typically thought of as having a privileged academic background.

There are several specific classroom variables more important than race or poverty, however, which may determine if another setting might benefit from a similar course of instruction. Math ability at the algebra I level or above is necessary to create relationships between columns in a spreadsheet, as these are equations in two or more variables. Any student failing to demonstrate this ability in this study would have received additional remediation, but this remediation was not necessary. Familiarity with computer spreadsheets, specifically graphing and manipulating data columns with equations (the students in this study were assumed to have this knowledge through completing Math 102, a required math class at this institution which uses Excel or equivalent) was required to construct the final model. Again, any student failing to demonstrate this ability would have received additional remediation as appropriate prior to the final modeling project. Finally, this curriculum was situated in a student-centered, inquiry-based, constructivist mindset. This feature was probably the most important feature of the classroom, as students and teachers often do not quickly adapt to a change from traditional

instruction to a less structured approach. Students may (and did) quickly revert to seeking the right answer; however in these lessons, there was not a singular right answer. Levels of frustration can (and did) rise very quickly. If a classroom was particularly teacher-centered and lecture based, with verification labs or no labs at all, these students would likely have a very difficult time with the constructivist approach, at least until they became comfortable with it. They would likely require additional supports in inquiry and active learning before starting the modeling unit.

Limitations

There are a number of threats to the internal validity of this study. The convenience sample chosen (four small classes taught by the author) presents a number of challenges. The small sample size (total $n < 60$) may not be large enough for some statistical tests. As these are the author's classes, there were potential questions of bias (which methodology will help to minimize) and successful application of this curriculum in situations other than when the author is teaching (again, see methodology). The threats and methods to address these threats are summarized below.

An independent scorer was used to score the free response questions on SUMS and SUSSI. An independent scorer familiar with the study but not part of it will provide an external check against potential bias the author may have when scoring the free response items on these two instruments.

A subset of the pretest was rescored mixed together with the posttests (double blind) to ensure that any tests receiving a higher score earned it on the merit of their answers, not from the assumption that students' answers would be better on the posttest.

Every attempt to triangulate all data was made between the subjective qualitative information collected (interviews, written assignments and drafts of the final modeling project) the subjective but quantitative (through rubric grading) free-response questions on the SUSSI and the SUMS and the objective Likert-scale questions on the SUMS and SUSSI. Conclusions not supported by all three areas were discussed.

All lessons pertaining to models were videotaped. It is the author's intention that all gains in nature of science understanding stem from the fact that the students apply what they have learned in refining models to their ideas of the nature of science, particularly regarding theory development. Gains in that area of the SUSSI could be achieved through direct instruction/memorization of these ideas. The videotaping of these lessons was to verify that this direct instruction has not occurred.

Whenever possible, all student-instructor interactions were audio taped.

Instrumentation threats should be minimal. The pilot study and revisions enacted during and after the pilot should allow for the instrument to remain the same.

Selection remains a risk. As students can choose not to participate in the study, or whether or not to participate in the interview, and can even choose to register for this particular class as their Area 10 (where word of mouth has already established that difficult mathematical modeling was present), some selection has already taken place. In the pilot study, students did not remove themselves from the study, and may not have given an honest effort on the assessments deemed not necessary to the class (by virtue of a grade). Increased attention to integration of all assignments including grades that reflect an honest effort was used to minimize this effect.

Other risks differ between the two course offerings. The brief nature of intervention (four weeks during the summer) minimized the threat of maturation during summer semester, although the 15 week fall semester had a greater chance for maturation risk. On the other hand, the longer class days during summer semester meant that each absence during this short period had a greater impact on the threat of history.

CHAPTER TWO

REVIEW OF THE LITERATURE

A review of the literature reveals information about the nature of the independent variable (cognitive developmental level and its measurement) and the dependent variables (understanding of the nature of science and conception of models and the measurement of each). Each of these areas was explored in the first section. There have also been studies that have attempted to improve each of these variables (cognitive development, understanding of the nature of science and conception of models) individually. Each of these areas was briefly addressed in the second section, with more emphasis placed on strategies involving teaching of and through models, because models were the medium through which the class was taught. The third section will discuss studies that have attempted to link the variables. No study in the literature has attempted to link all three, thus, the potential importance of this study. Where appropriate after each section, key studies will be analyzed more critically.

Independent and Dependent Variables and Their Measurement

Cognitive development.

Any discussion regarding cognitive development must begin with Jean Piaget, as the idea of developmental stages began with his work. In brief, children, as they develop, pass through a series of stages from birth to adulthood. The order of the stages is invariant, although initially Piaget had tighter age brackets for when these transitions typically occurred than is now observed. These stages are sensory motor, typically from birth through 24 months, pre-

operational (or conceptual or socialized thought) from age two to seven, concrete operational thought from approximately age seven to age 12, and full formal operational thought beginning as early as age 11 or 12 (Piaget & Inhelder, 1955). Piaget and Inhelder (1955; 1966) describe several key transformations that occur between the concrete and formal stages. The first is the ability to reason using hypotheses, and subject these hypotheses to testing and verification. This transition is of central importance to tasks in this study since hypothesis testing is central to the nature of science, models, and cognitive development. Tasks in this study will require the student to isolate and systematically change independent variables to determine the extent to which this results in a change in the dependent variables, a skill which is consistent with abilities first seen at the formal operations stage. The transition from concrete to formal operational though also includes using propositional logic, symbolic logic, and combinatorial thinking, however, these processes are less directly important to this study.

Critiques of Piaget's works have found one central flaw in his work that is related to this study. In many settings the onset of formal operational thought may not appear until much later than age 11 or 12, if at all. Lawson, Clark, Cramer-Meldrum, Falconer, Sequist and Kwon (2000) found that 45% of their sample of college students were unable to reason formally on a consistent basis (11% concrete operational and 34% transitional formal operational), even though the mean age of the students in this class was 20 $^{+}/_{-}$ 3 years, well beyond the age at which formal operational thought has the potential to be developed. This development, or lack thereof, appears to be linked to the experiences of the developing child during the appropriate time for growth. A central tenet of Piaget's theory of development of cognitive stages is that a student must be presented with stimuli at the next level of development, and it is this stimulus that causes a rearrangement of the student's thinking. It would seem that college students in an

industrialized nation should certainly have achieved the formal operational level of development, but increasingly this is not the case. It is possible that students are not succeeding at cognitively demanding tasks that require formal operational thinking, if they are still in concrete operational thought or just transitioning from concrete operations. If such a difference in cognitive developmental level exists, it is necessary to measure this level accurately to determine its effect on classroom learning.

Another stage of development appears in some people after age 18 and is a post-formal stage, or as Lawson, Banks and Logvin (2007) call it, the fifth stage. At this stage, students are able to reason hypothetically about processes that involve unseen agents. Take for example, atoms and molecules. As the atoms and molecules about which the students in this study are postulating are unseen agents, this level of reasoning would seem to be helpful. However, other considerations make this relationship less clear. First, these ideas of atoms and molecules have become so familiar and ubiquitous that, even though they are abstract constructs, they have even entered the everyday language, such as H_2O for water (Harrison, 1998). Harrison points out that this familiarity may cause both students and teachers alike to view these atoms as facts rather than constructs. This idea that the abstract can become more concrete is at odds with statements to the contrary in Lawson, Clark, Cramer-Meldrum et al. (2000); even the electron microscope that allows atoms to be "seen" does not move atoms from the realm of theoretical to concrete, since the image only shows "little round balls" (p. 85) and it is still up to the individual to ascribe theoretical properties to these balls. Second, once numbers (concentration of carbon dioxide measured at Mona Loa, for instance) associated with these constructs are available, the students may be able to manipulate the numbers successfully without a solid grasp of the underlying chemical concept, from a purely mathematical standpoint. Third, it may be possible for students

in some situations to perform apparently above their conceptual level by rote. Adey and Shayer (1990) criticize this in the literature as "training," since the student learns to do a task above his cognitive level, but has no ability to transfer localized conceptual development to other situations. Lawson, Alkoury, Benford et al. (2000) consistently show students classified at the concrete level who occasionally succeed on formal and post-formal tasks, especially if students have a high level of declarative knowledge, (see also Lawson, Clark, Cramer-Meldrum et al. (2000); Lawson, Drake, Johnson et al. (2000)).

Nature of science.

The nature of science (NOS) is a broad field. Abd-El-Khalick, Bell and Lederman (1998) clarify the field somewhat by separating the philosophy of science and the nature of reality from an operating knowledge of NOS that would be expected of and useful to K-12 students. This operating knowledge includes the key NOS ideas that "scientific knowledge is tentative … empirically based … subjective … the product of human inference, imagination, and creativity … and socially embedded" (p. 418). Understanding the differences and relationships between observation and inference and between theories and laws is also important.

One particular misconception regarding the nature of science discussed in Windschitl and Thompson (2006) is the oversimplified idea of a hypothesis that may result in content-free inquiry. Because of the simplified version of the scientific method presented by cookbook labs, verification labs, and science fair projects, student are confused about the nature and relationships between hypotheses, laws, and theories. Windschitl and Thompson (2006) point out that students do not see theories as being central to generating new hypotheses, despite the fact that the *educated* in *educated guess* in the definition of scientific hypothesis means grounded in theory or scientific model.

In addition, many students subscribe to a hierarchical view of hypothesis, theories and laws, specifically that hypotheses become theories if they are right and eventually become laws with further testing. This misconception could be related to the fact that students are not aware that there are different types of hypotheses. A hypothesis that variable *a* is correlated to variable *b* could, if well supported by data, lead to a law. For example Charles' Law states that volume of a gas changes directly with absolute temperature. A hypothesis that attempts to explain a phenomenon is more likely to eventually generate a theory. For instance, the Kinetic Molecular Theory hypothesizes that a gas is made of little invisible particles (atoms and molecules) that move in random motion, taking up almost no space and with almost no attraction to each other, that pressure is caused by the collisions these molecules make with each other and the walls on the container and that temperature is proportional to the average kinetic energy (energy of motion) of these particles. From these propositions, the following chain of logic can be made. The increase in temperature causes an increase in kinetic energy of the particles which causes the molecules to move faster and thus have more collisions with each other and with the walls which results in increase in pressure against the walls, which would tend to force them out and expand the volume of the container. The hypothesis of the kinetic molecular theory gave an explanation of not only Charles' Law, but each of the other gas laws as well. Since the hypothesis generating a law merely looks at a correlation between two directly observable phenomena whereas the theory generating hypothesis postulates the existence of other variables that explain not only the initial relationship being studied, but several other relationships, it should be obvious that the misconception that hypotheses become theories which then become laws is flawed, yet, since students lack authentic experience creating hypotheses this misconception continues to exist.

Several tests exist for measuring understanding of the nature of science. Lederman, Wade and Bell (1998) documented over 20 instruments in the 40 years prior to their study. The Views on the Nature of Science questionnaire (VNOS), created by Abd-El-Khalick, Lederman, Bell, and Schwartz (2001) was one of the most widespread instruments for measuring NOS in the United States, whereas the Views on Science-Technology-Society (VOSTS) created by Aikenhead and Ryan (1992) had widespread usage in Canada,. These two tests approach the topic very differently, with the VNOS consisting of a few (10 or less, depending on the version) open-ended questions followed by semi-structured interviews for clarification. Its authors argue that trends noted in other tests, particularly those employing forced-response questions, reflect the ideas of the test writer more than those of the students taking the test (Adb-El-Khalick, Bell, & Lederman, 1998). The VOSTS, on the other hand, is a massive 113 question, empirically-derived, multiple-choice test that, in addition to measuring NOS, also considers heavily the relationship of science to technology and society, as the name implies. Aikenhead and Ryan (1992) contend that ambiguity is still present in free-response NOS instruments (such as the VNOS) at a high level, unless the time-consuming follow-up interviews are performed; however, by using empirically-derived multiple-choice tests, the ambiguity can be reduced to levels (15%-20%) only slightly above levels found in clinical interviews (5%), and certainly better than the 35%-50% ambiguity reported for paragraph answers (Aikenhead & Ryan, 1992). Thus, it is apparent that eventually a compromise between these two extremes might be presented.

Liang, Chen, Chen, Kaya, Adams, Macklin, and Ebenezer (2006) have produced a test that attempts to combine some of the best aspects of both the VNOS and the VOSTS. One criticism of the VNOS and other-open ended tests in general is that there is potential for students to not give their best guess and instead leave an item blank. This non-answer hampers the ability

to interpret students' understanding because it is not known if the student knows the correct answer, but is unsure and wishes not to guess, has absolutely no knowledge of the topic in question, or has decided not to answer for other reasons such as test fatigue or obstinacy. By combining a set of Likert-scale questions with a free response question within each of the key NOS areas (hereafter referred to as sub-scales) mentioned previously, the Student Understanding of Science and Scientific Inquiry questionnaire (SUSSI) developed by Liang et al. (2006) strikes a balance between the reliable, easy scoring and almost guaranteed answers from the Likert-scale questions with the opportunity for students to give more complete and detailed answers to the free response. Disparity between the two halves (Likert-scale and free-response) of each sub-scale could be used to identify questionable data (if the Likert-scale questions are correct but nothing is written in the free-response, then it is more likely that this student chose not to answer than that this student had no knowledge of the subject). The SUSSI has undergone multiple validity tests with science experts including scientists, science educators, and historians and philosophers of science, and with three samples ($n > 200$) of pre-service science educators in the U.S., China, and Turkey. Cronbach's alpha is 0.67 in the American and Turkish sample, but only 0.61 in the Chinese sample, indicating some degree of reliability.

One troubling comment in both Liang et al (2006) and Aikenhead and Ryan (1992) is that the conventional concepts of validity and reliability may not apply to an empirically derived instrument. Aikenhead and Ryan (1992), citing primarily the work of Mishler, (1990) claim that validity rests in trust that one researcher has for another. Furthermore, Abd-El-Khalick et al. (2001) and Liang et al. (2006) rely heavily on construct validity, reasoning that if the science experts (those with doctorates in science, science education, or a related field) score higher than novices, the instrument has construct validity. Abd-El-Khalick et al. (2001) used science novices

of the same age and educational level (doctorates in a non-science field) as their science experts and demonstrated this difference. Liang et al. (2006) did not have a comparable group of novices.

Modeling.

The National Committee on Science Education Standards and Assessment (NCSESA) defines a model as "tentative schemes or structures that correspond to real objects, events, or classes of events, and that have explanatory power. … Models take many forms, including physical objects, plans, mental constructs, mathematical equations, and computer simulations" (NCSESA, 1996, p. 117). Modeling, therefore, is the process by which a model of a phenomenon is constructed.

Exploratory, expressive, and explanatory are three words sometimes used to describe models, and differentiate based on who creates the model and how the model is used. Mellar and Bliss (1994) define exploratory models or modeling as using someone else's assumptions and expressive models or modeling as expressing one's own ideas and assumptions. The second type of modeling appears, at least at first, to be more consistent with constructivist ideas about learning than the former. Typically when students use models to understand a phenomenon, they use models created by others such as textbook authors and teachers. An explanatory model is a teaching model, used by an instructor, to make a concept more accessible to the student (Clement, 2000; Clement & Steinberg, 2002). While the student may work with manipulating the explanatory model in an exploratory way, the students themselves are not actively modeling the phenomenon; they are not making decisions about which data or variable to include in the model and how these ideas should be related. Thus, in order to truly understand models and modeling, using models created by others alone is not enough if one espouses this view. Others

(Justi & Gilbert, 2002a; Lesh & Doerr, 2003) see student manipulation of an explanatory model as an essential step in learning to model. This manipulation can rival the learning found in the creation of an exploratory model, since to effectively manipulate the model, a student must intimately understand its construction.

Many phenomena are best approached with multiple models, yet students have little conception of the purpose of multiple models of the same phenomenon. For instance, models of the atom include the octet rule, the Bohr or planetary model, space filling models, sticky balls model, Lewis Dot Structures and some others with further refinements. Electric circuits are often compared both to dams on a river (series circuits) and cash registers at a supermarket (parallel circuit). Each model has particular strengths and weaknesses for explaining different behaviors of atoms or circuits. One particular difficulty that students experience is the complexity of having multiple models for the same phenomenon. Part of this difficulty stems from students' fundamental misunderstanding of the purpose of a model. No one analogy (or model) can capture the entirety of a phenomenon (Clement & Steinberg, 2002). If a student feels that a model is a copy of reality, then models can only be evaluated on how well the model matches reality.

Another difficulty that students have with models relates to the expert vs. novice issue seen elsewhere in science education. An expert is familiar with the purpose of each model and is able to select the model that is best suited to the application, often the simplest model that accurately predicts the particular behavior of interest. A student, however, may either look for the one best model to use indiscriminately or select the wrong model for the application. Selecting the wrong model may occur because of misidentification of the particular aspect being studied, based on surface feature analysis, rather than understanding the concepts involved in the

problem and comparing those concepts to the strengths and limitations of each model. The selection of approaches based upon surface features has been well studied in problem solving in physics, is commonly seen in novices and rarely in experts in the field (Chi, Feltovich, & Glaser, 1981; Schoenfeld, 1982). Lesh, Cramer, Doerr, Post, and Zawojewski (2003) state that this focus on surface features comes from poorly integrated knowledge, but that presenting multiple perspectives routinely during class can help. This idea is further supported by Gutwill, Frederiksen, and White (1999) who found that students presented with an integrated model scored lower on content tests than students presented with multiple, discrete models of electrical current. They hypothesized that students were forced to construct their own meaning from the multiple models, as well as reconcile differences between these models, while the students presented the integrated model were able to be more passive in their learning.

Many of the problems that students have with models may stem from the way these models are taught in science classes. Justi and Gilbert (2002a) emphasize that teachers do not spend enough time discussing the scope and limitations of each model, a practice that increases the chance that a student may select the wrong model for the situation. It is essential that a student has a solid "anchoring conception" (Clement & Steinberg, 2002, p. 403) before mapping can begin from the model to reality. An anchoring conception is defined as "useable working knowledge … that can be used as the basis for an analogy" (Clement & Steinberg, 2002, p. 403).

Another misconception about multiple models seems to come from a recent trend in education theory. This misconception is shared by students (Chittleborough, Treagust, Mocerino & Thapelo, 2005) and teachers (Cullin & Crawford, 2003) and is that multiple models relate to the concept of learning styles. For instance, one model might be better for a student with a visual learning style, another for a student with a kinesthetic learning style. However, models of the

same situation or phenomenon have more to do with the ability of a particular model to better explain one aspect of a particular concept than accommodating the individual learning styles of students. An example might clarify. Molecular model sets (used in organic chemistry for covalently bonded molecules to help them visualize three-dimensional shapes and bond angles) are plastic balls (which stand for atoms) and sticks (which stand for bonds) that are physically manipulated, and thus consistent with kinesthetic learning. A Lewis Dot Structure is pencil and paper model of the atom showing its outer shell electrons, and is less kinesthetic. However, although both models can describe covalent bonding, the molecular model set cannot be used for ionic bonding. Therefore, it is the purpose of the model, in this case the three dimensional structure of the covalent molecule versus the flexibility of being able to show both covalent or ionic bonding, that decides which of the two models should be used to show ionic bonding, not the learning style of the user.

Students typically do not perceive models was science educators would like them to. The National Committee on Science Education Standards and Assessment (NCSESA) (1996) states that middle and high school students tend to see models as physical copies of reality, rather than as representations of ideas. Chittleborough, Treagust, Mocerina, & Thapelo (2005) find approximately one quarter of their grade eight through ten students selected "accurate duplicate of reality" as a definition for model rather than "a representation." When the same survey was given to students in higher grades, the percentage choosing "accurate duplicate of reality" decreased. All percentages were substantially lower than Grosslight et al. (1991) found, where nearly 50% of students conceived of models as duplicates of reality. Although many models (and most models students have experience with outside of class) take a physical form, in science the word model is just as likely to be synonymous with *hypothesis*, *law*, and *theory* (NCSESA,

1996). This predictive and explanatory power of models is found to be underappreciated by students (Chittleborough et al., 2005).

Students' lack of familiarity with the non-physical meaning of the word *model* does not prevent students from carrying out pattern seeking on a daily basis. Student misconceptions are typically based on a mental model of how a phenomenon works that is often at odds with the scientific model (Duit & Glynn, 1996 in Chittleborough et al., 2005). Because humans are by nature pattern-finding animals, students can and will construct their own meaning from observations they make in everyday life. Students who are more comfortable with science being a dynamic field (Songer & Linn, 1991) and are exposed to a curriculum where students are allowed to build and revise their own theories (Carey et al., 1989) have better ability to integrate new knowledge into their existing framework and a better understanding of the nature of science.

Like students, teachers too have misconceptions concerning models and modeling. Although the NCSESA (1996) states that it is a responsibility of teachers to move students toward a more scientific understanding of models, teachers whose own understanding of models is tentative may find difficulty in changing student perceptions. This uncertainty, coupled with a focus on content instead of the process of science, prevents gains from being made in student appreciation of models (Justi & Gilbert, 2002b).

Practicing scientists perceive of models differently than either students or teachers. Students need to understand models in order to learn science, to learn about science, and to learn how to do science (Justi & Gilbert, 2002a). They state that learning how to model is centered on the student's mental model, and is a process that involves understanding others' models, the student revising existing models, and making his or her own models (Justi & Gilbert, 2002a). A mental model is the understanding that one has about how a phenomenon behaves, and may or

may not be consistent with accepted scientific models. Lehrer and Schauble (2003) state that scientists and mathematicians see their fields as a "process of constructing, investigating, applying, interpreting, and evaluating models" (p. 59). Schwartz and Lederman (2005) conduct a survey of 24 practicing research scientists, averaging 25 years of research experience across many fields, who give their views on models. These scientists also consistently mention the themes of using models to make and test hypotheses, but also talked about models as important for organizing observations, and over one third discuss the mathematization of a system.

The process of modeling, or model building, appears to fall outside of the student conceptions of models listed above. As much modeling is mathematical in nature, and since students perceive models as tangible objects, students do not see most modeling as being related to their conception of models. However, Ogborn (1994) counters that students grow up representing reality with boxes and sticks (which will not be exact copies) as much as they do with dolls and toy trains (which are closer to scale models). Therefore, it should not be assumed that students are unready to model. Later games, such as Monopoly, are based on a model of the economy. However, since students may not be consciously aware that they have been modeling all their life, they may have difficulty attempting to model for the first time in a classroom setting.

Lesh and Doerr (2003) define a *model-eliciting activity* as involving "sharable, manipulatable, modifiable, and reusable conceptual tools (e.g., models) for constructing, describing, explaining, manipulating, predicting, or controlling mathematically significant systems" (p. 3). Students construct these models while solving authentic problems. Many of these concepts (manipulatable, modifiable, and reusable, for instance) have strong crossover with theories.

A mathematical model is a mathematical relationship such as an equation or series of equations, which show the relationships between variables in a system. While typical mathematical models use quantitative relationships, semi-quantitative or even qualitative relationships may be used. Mathematical models may then be classified by these relationships; a model with qualitative relationships is a qualitative model. Typically, these less quantitative models sacrifice accuracy for simplicity, and are therefore more appropriate for students learning about models or learning to model, and are less appropriate when trying to use a model to answer a question. However, an important use of qualitative and semi-quantitative models is as an intermediate step in constructing a quantitative model.

Quantitative modeling is a sub-category of mathematical modeling, and is the most common form of modeling. In quantitative modeling, algebraic relationships are established between the phenomenon (variable) to be studied and the other variables which might influence that phenomenon (variable).

On the other hand, true qualitative modeling rarely exists by itself. There are rarely situations where stating *a* increases as *b* increases is sufficient, without further clarifying the relationship. However, Harrison (1998) argues that students should generate qualitative understandings of quantitative models, and many researchers (Justi & Gilbert, 2002a and others, detailed below) see qualitative understanding as a step on the path to quantitative modeling.

Although purely qualitative modeling is rarely an end onto itself, semi-quantitative can be. Semi-quantitative modeling includes determining the relationship between variables. These relationships are limited to more general relationships, as opposed to algebraic functions: as X increases, Y either increases, remains relatively constant, or decreases. If desired, the next step is to examine the rate of increase or decrease in the relationship: as X changes a little, Y changes

a little or a lot. Finally, the concavity can be examined: as X changes, Y changes more and more, less and less, or about the same each time. By changing these inputs, modelers can generate graphs relating the input variables and the response without actually knowing the appropriate algebraic function. This type of modeling can be accomplished with the aid of modeling software. *ModelsCreator* and *Model-It* are programs that create models by working from a semi-quantitative relationship (Ergazaki, Komis, & Zogza, 2005). Students first examine the agents or "entities" that contribute to a model, then determine the specific variables or "properties" associated with these agents, and finally discover the relationship between these variables. These relationships are explored visually, using graphs or charts, so that students are not required to create algebraic expressions (Ergazaki, Komis, & Zogza, 2005, p. 911). The authors also feel that the use of visuals to explore relationships allows students to become more metacognitive about model building (Ergazaki, Komis, & Zogza, 2005).

An example of each type of relationship can be seen in Figure 1, as pertains to height and mass data from the Center for Disease Control (CDC, 2007).

Qualitative relationship. As the height of a boy increases, his mass increases.

Semi-quantitative relationship. As the height of a boy increases, his mass increases faster and faster.

Quantitative relationship. The equation that describes the relationship between height and mass of a boy is mass = 0.0061*height1.707. This equation would also represent a quantitative model.

Mathematical models are constructed using three strategies: specific modeling software, spreadsheets, or computer programming. Semi-quantitative modeling is identified more with modeling software, and quantitative modeling with spreadsheets or computers. Each method has

Height (cm)	P50 Mass (kg)
77.5	10.38902
80.5	11.07049
83.5	11.74598
86.5	12.42280
89.5	13.10893
92.5	13.81254
95.5	14.54147
98.5	15.30286
101.5	16.10277
104.5	16.94631
107.5	17.83808
110.5	18.78316
113.5	19.78744
116.5	20.85633
119.5	21.99204

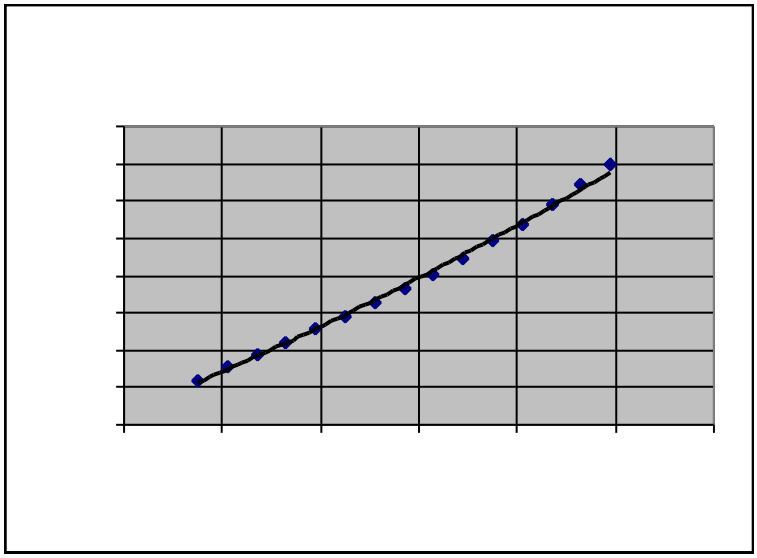


Figure 1. Height vs. mass data for boys

its advantages and disadvantages, but computer programming requires specialized training and is beyond the scope of a liberal studies course. The other two methods are compared below.

Model-it, Agent Sheets, NetLogo, DMS, STELLA, CMS, Model Builder, Algebraic Proposer and other software have been developed to help students model. Two advantages of

these programs over spreadsheets are that many of these programs allow students to make semi-quantitative models, and the use of objects and icons make them far more visual. Both of these advantages are seen as particularly important to the younger modeler.

These modeling packages also have several disadvantages when compared to spreadsheets. First, many are not free. Second, since students are unlikely to have previous experience with the software, the learning curve is steeper. Third, many of these packages appear to support only correlational formulas, not recursive formulas. In other words, A can depend on X, Y, and Z, but A_1 cannot depend on A_0. Recursive formulas are a powerful way to model some phenomena (like exponential growth and decay) and so this limitation is non-trivial.

If specific computer programs are not used to make quantitative and semi-quantitative models, spreadsheets are the other main option. Spreadsheets have several advantages over dedicated modeling software: they are commonly available, they are relatively easy to use, and many people know how to use spreadsheets already so little training is needed, unlike dedicated modeling software. In addition, spreadsheets are versatile and, depending on the spreadsheet, they can be used to create a variety of models such as static models, "what ifs?" trial and error models, and even iterative dynamic models and probabilistic models (Boohan, 1994). Disadvantages of spreadsheets compared to dedicated modeling software include the fact that they are typically limited visually, users lacking algebra skills may have difficulty representing links between variables (as actual equations must be written), and relationships are typically defined in terms of cell or column references, instead of meaningful names or variables. When iterative models are constructed, the necessary changing of the range of cells when changing the formula can be time consuming. While spreadsheets can graph, the graphical displays provided can be inconvenient to use (Boohan, 1994), although they have improved.

The previous example using height and weight shows a natural progression from the simple to the complex in a mathematical model. The same progression can occur in a non-mathematical case. Clement (2000) notes that students need a spatial model of layers of the earth (a simple model) before they can have a conceptual model of plate tectonics (a more complex model). Lesh and Doerr (2003) feel that this progression is a natural part of modeling as explored further below.

Having established what modeling knowledge is, the next step is to determine appropriate instruments and procedures for its measurement. Many studies that have attempted to measure modeling knowledge have relied on an interview procedure. Grosslight et al. (1991) established a classification or scoring system for modeling ability measured on a three-point scale, with those seeing models as copies of reality with the purpose to show something visually at the first level, those seeing models as differing from reality in some ways, with some aspects emphasized and others simplified to aid in communication at the second level, and those seeing the purpose of models as generating testable hypotheses at the third and highest level. Many modeling studies that followed use this same classification scheme and basic interview protocol. Many modeling studies have adopted this more or less standard classification of student modeling ability, or at least have been influenced by them. A counterpoint to this article would be Schwartz and Lederman (2005) article mentioned above, which does perhaps give a slightly broader expert view of models, but never-the-less does not conflict with these broad classifications on the major purpose of models.

Although interviews may be the most accurate method, a need exists for an assessment of modeling that may be administered to a large group of students at the same time. One such instrument, the Student Understanding of Models in Science (SUMS) instrument, was created by

Treagust, Chittleborough, and Mamiala (2002). This instrument has high internal reliability (0.71 to 0.84), and because it is a Likert-scale instrument, requires no interpretation on the scorer's part, no multiple graders, nor any inter-rater reliability measures. However, it fails to address the question of how models are created, the role of the modeler, and it gives less emphasis to the role of models in generating and testing hypotheses than Grosslight et al. (1991) did.

Critical analysis of key studies.

As many of the modeling studies refer back to the Grosslight et al. (1991) paper, the analysis should begin with this paper. This study references earlier anecdotal reports regarding modeling knowledge of students, but purports to make a significant step forward in the assessment of modeling knowledge through the use of clinical interviews and obtaining an responses at a variety of levels (seventh grade, 11th grade, and expert). This variety of responses was expected to provide a continuum for judging the modeling knowledge of students, since new modeling curriculums, like the nature of science, was a topic of great interest at that time.

The student samples (33, mixed-ability seventh grade students from suburban Boston, 27 honors 11th grade students from suburban Boston) should provide the claimed variety of abilities. A possible criticism of methodology comes from selection of the expert group. The expert group consists of a science museum director, a high school physics teacher, a professor of education and engineering, and a researcher in thinking and representation. Of the four experts used to determine the expert view on scientific models does not appear to contain a single practicing scientist. There appear to be no working Ph.D. physicists, biologists, chemists, or geologists. Particularly since Grosslight et al. criticize a similar previous study on the nature of science for not gathering data from experts, the lack of scientists in their study could have been a serious

blow. However, their results generally concur with other consensus views on the nature of scientific models. Furthermore, a study by Schwartz and Lederman (2005) provided a more comprehensive sample 24 veteran research scientists, who generally support the findings of Grosslight et al. (1991)

The questions used and the format of the semi-structured interview appear appropriate, with the exception that the model prompts did not include a mathematical model (the prompts were a toy airplane, a subway map, a picture of a house, and a schematic diagram). Did the lack of a mathematical model skew their data away from mathematical models? Only three percent of 7^{th} graders and 14% of 11^{th} graders identified mathematical models as an option.

A second key study regarding the assessment of modeling is Treagust, Chittleborough and Mamiala (2002), which produced the SUMS instrument used in this study. This 27 question, Likert-scale survey aims to assess several aspects of modeling knowledge with separate sub-scales. These sub-scales are: (a) multiple representations of scientific models; (b) whether or not scientific models are exact replicas; (c) the explanatory nature of models; (d) the use of scientific models; and (e) if, how, and why scientific models change (Treagust et al., 2002). This study too had a cross-age sample (ages 13-15) that could help to define levels within the instrument. This study, however, lacked the expert views to define the highest conceptions on the scale.

It is the construction of the test, however, where the most serious flaws in the methodology occur. First, for an instrument called Student Understanding of Models in Science, it leaves some of the most important characteristics of models un-assessed. One aspect of modeling that is not addressed by this test is the construction of models themselves. While the *if, how, and why scientific models change* sub-scale measures ideas related to the modification of models, which ought to be similar to ideas concerning construction of models, this aspect of the

modeling is significant enough that it deserves its own questions. Three out of seven of characteristics of models described in Van Driel and Verloop (1999) involve model design including (a) some aspects of the phenomenon intentionally excluded from the model for simplicity's sake (b) conscious choice in the selection of factors to include and exclude, and (c) development through an iterative process including comparison to empirical data. Justi and Gilbert (2002a) also mention observations and data sources as the start of model building, and the importance of testing with empirical data. Perhaps the key aspect of a scientific model is its ability to generate a testable hypothesis (Van Driel & Verloop, 1999).

There was no validity measurement reported for this instrument, as with the NOS instruments, perhaps because no established, valid criteria or scale exists for measuring modeling ability exists. Cronbach's alpha for each of these sub-scales ranged from 0.71 to 0.84, indicating a high degree of self-consistency and thus reliability of the data. Item to total correlations were above .45 for 26 of the 27 items, and a bi-variate correlation of each sub-scale was significant at the .01 level. While these statistics at first look promising, a closer examination revealed an alternative explanation for this high correlation. Almost all questions were phrased to the positive, such that the most correct response, the most likely response (in the case of misconceptions) or both was "strongly agree." This is perhaps why there is such a high correlations between sub-scores; as most respondents answered "agree" to most answers.

Conclusion.

There is some similarity between the concepts of modeling, nature of science and cognitive developmental and their assessment, and these will be discussed further in the third section of the literature review. Each is best assessed with clinical interview protocols. Only cognitive developmental level seems to have found a valid and reliable pencil and paper test in

Lawson's CSTR. The nature of science and modeling assessments appear to be still settling on an accepted format.

Improving Nature of Science Knowledge, Modeling Knowledge, and Cognitive Developmental Level

Cognitive development.

Adey and Shayer's work is central to understanding how a student's cognitive development can be accelerated and thus the student's developmental level can be increased. In their massive, four-part study across multiple schools (Adey & Shayer, 1990; Shayer & Adey 1992a, 1992b, 1993), Shayer and Adey showed that students who are continually exposed to problems and activities in science requiring formal reasoning achieved great gains in their developmental level (pretest to posttest). Furthermore, these gains in developmental level were permanent and translated into gains in multiple areas. Not only were the students exposed to this program as 11-12 year olds better able to reason formally, but at age 16, they dramatically outscored their peers on national standardized exams in science and mathematics, and surprisingly English as well. This gain in science and mathematics knowledge could be attributed to better science and math instruction during the intervention. However, the improvement on the English portion of the exam, which in no way should have been directly influenced by teaching of formal reasoning in a science class, points to increased global cognitive development by the program. In lay terms, the kids got smarter, not just knew more

Nature of science.

Much attention has been paid to the history and nature of science in early twenty-first century curriculum reforms. However, this is not the first time that the nature of science itself

has been addressed as an area worthy of instruction. Here, the work of Abd-El-Khalick and his critical review of the literature are important.

Despite emphasis on inquiry and hands-on learning in the 1960s and 1970s, students did not learn science merely by doing science (Abd-El-Khalick, Bell, & Lederman, 1998). Thus, something more than being exposed to scientific inquiry was necessary. Akerson, Abd-El-Khalick and Lederman (2000) studied pre-service Master of Arts candidates in secondary science education and found that an explicit reflective approach to teaching NOS yielded raw gains in students' NOS understanding of 8% to 56%; with typical gains being approximately 20% raw and 33% normalized gain. Therefore, it is not just exposing students to experiences in science that should make them better at understanding NOS. It is when these experiences are accompanied by activities to make students focus on how their NOS conceptions have been changed by the experiences that is important.

Modeling.

Because modeling is seen as a central idea to science education, various mathematics and science educators have attempted to teach modeling through a variety of strategies. Most involve incremental procedures that build modeling knowledge. Many seem to relate well to the Learning Cycle, and some seem to mirror Piagetian cognitive development in miniature.

As discussed previously, students rarely model in the scientific sense, but some common classroom activities, such as creating concept maps, can be early steps in the modeling process. When students create models, however, they typically start with form over function (Lehrer & Schauble, 2003). This tendency seems to be related to their disproportionate exposure to physical models, and also to any informal modeling they may have done at play. Their initial models start by looking like what they are modeling and proceed to making the model behave

like the target (Lehrer & Schauble, 2003). However, in the case of a mathematical model, this physical resemblance is difficult to achieve. For many environmental issues, a cycle such as the carbon cycle (itself a model) might provide that initial form that allows students to proceed on to function. In the event that such an existing conceptual model is not available, the construction of a concept map (qualitative model) relating the ideas to be modeled seems to be a common first step.

One group using such an approach is Ergazaki, Komis, and Zogza (2005). Ergazaki et al. use dedicated modeling software (called *ModelsCreator)* to help students construct their models. The software allows students to build a qualitative, graphical relationship between variables, and then manipulate these relationships in a semi-quantitative manner. One strength of *ModelsCreator* is its ability to create an output log, which records the changes made to the model as it evolves. Particularly from a research standpoint, this approach allows teachers and researchers to not only see the finished model, but also the modeling process.

Before even beginning the first modeling cycle, Lesh, Cramer, Doerr, Post, and Zawojewski (2003) recommend that "warm-up activities" (p. 45), such as mathematizing a newspaper article or other written work should take place in order to prepare students for modeling. Mathematizing means to take the ideas and relationships expressed verbally in the paper and translate them to variables and equations, and is often the most difficult task for students (Lesh & Doerr, 2003) because it is not a skill traditionally taught in the curriculum. When learning about a concept, three distinct types of modeling activities are used. First it is recommended that students complete one or more class periods of model-eliciting activities. This model-eliciting experience could begin as a whole group activity and progress to small groups. Students should practice looking for relationships, mathematizing the information, and

working with the numbers. These model-eliciting experiences are followed by model-exploration activities, where the various related models are used to strengthen the overall understanding of the phenomena. Finally, model-adaptation/application/extension activities try to use the tool developed within a fairly constrained context to attempt to solve a problem outside of this context. These activities are followed with reflection, debriefing, and follow-up activities that are designed to make the student metacognitive about the modeling process and to reinforce the learning, which is similar to the explicit-reflective process of Akerson, Abd-El-Khalick and Lederman (2000) for teaching NOS.

Both Lesh, Cramer, et al. (2003) and Justi and Gilbert (2002a) discuss several underlying assumptions and principles that should influence construction of modeling lessons and curricula. Before delving into the specifics of how a specific modeling activity should be structured, an examination of these principles would be illustrative.

Lesh, Cramer, et al. (2003), approaching modeling from a mathematics standpoint, show how Dienes' Instructional Principles (1960) relate to modeling instruction. These principles are: (a) *Construction* of a concept occurs from systems rather than from concrete objects through reflective abstraction, (b) In order to become metacognitive and think about a model instead of thinking with a model, students must see *multiple embodiments* of the model, (c) The system must be (and must seem to be) *dynamic*, with emerging patterns, rather than static, and (d) The multiple embodiments should have built in, insignificant differences (perceptual variability) that become filtered out when comparing the multiple embodiments (Lesh, Cramer, et al., 2003, p. 37-38). How these principles shape instruction would seem to indicate that when learning about and with models, more than one model must be used to achieve principles (b) and (d). Principles (a) and (c) require that careful consideration be made of the problem to be modeled, as a static

situation does not allow for modeling, and neither does modeling of physical objects. Mathematical models would appear to lend themselves well to all four principles.

Lesh, Cramer, et al. (2003) also present six principles of instructional design with models. These are (a) The Personal Meaningfulness Principle (reality principle), (b) The Model Construction Principle (looking for patterns, which cannot be solved as a rote exercise), (c) Self-Evaluation Principle, (will students be able to assess their model?), (d) Model Externalization Principle (will students have to explain their thinking?), (e) Simple Prototype Principle (is the situation as simple as possible, yet still allows for creation of a prototype model that can be used in other situations?), and (f) The Model Generalization Principle (can this model be modified to apply to other situations?).

These principles are important in shaping modeling instruction in much the same way Dienes' instructional principles were. Principle (a) (Personal Meaningfulness) requires a modeling topic that can be approached in a variety of ways. Principle (b) seems to point to a dynamic system with more than one answer, and principle (e) constrains the difficulty level of the modeling project, and thus, like Dienes' principles (a) and (c), requires the instructor to put much thought into the system assigned for study. Principle (f) has ties with theory building, in that a good scientific theory is useful for explaining a wide variety of phenomena. Principle (d) is similar to the qualitative concept maps used in other approaches to explain the reasoning. Finally, principle (c) relates to the ability of students to be able to make predictions (hypotheses) and test their models against reality (and then revise), without relying on an external authority to determine if their model is good or not.

Justi and Gilbert (2002a) present a slightly different, step-wise approach to teaching students to create models from scratch. Justi and Gilbert's focus is broader, as it is not

constrained by a specific piece of software, like Ergazaki et al., nor was it as focused on steps within creating a single model like Lesh and Doerr (2003). Instead, it appears to be more focused on the pedagogy of teaching students how to model over a series of successive lessons. Their five steps of teaching to model are (a) learning models, (b) learning to use models, (c) learning how to revise models, (d) learning to reconstruct models, and (e) learning to construct models de novo (p. 369). This is an incremental approach, with the students gradually increasing their role in the modeling process. Students first become familiar learning about models and using existing models, and then proceed on to minor and major revisions of models before ending with constructing their own models from scratch. These minor and major revisions of models seem to parallel the *evolutionary* and *revolutionary* changes that theories undergo (Liang et al., 2006).

White (1993) and Schwarz and White (2005) are both articles dealing with the ThinkerTools curriculum. While White (1993) is concerned primarily with mental models, i.e. the rules that students internalize and use to approach answering questions, Schwarz and White (2005) involve students constructing models (the object) as well as mental models (the idea). The first gives some key reasons for the success of the Thinker Tools approach in helping younger-than-average students to construct a strong mental model for understanding Newtonian Mechanics. Most relevant to this study is the idea that development of a model should proceed from simple to complex, with data presented in such a way that students can find the easy patterns first, and then incorporate the nuances over several revision cycles. Schwarz and White (2005) use goals similar to Justi and Gilbert (2002b) that students should learn (a) the nature of models, (b) how to create models, (c) the evaluation of models, and (d) the utility of modeling (p. 1289).

Several modeling cycles will be discussed. Figure 2 attempts to show how the various approaches to a modeling cycle overlap with a learning cycle approach to teaching.

Lesh and Doerr (2003) and Lesh, Cramer, Doerr, Post, and Zawojewski (2003) present many sets of principles (detailed previously) for structuring modeling instruction, from the overall organization, to topic selection, to the structure of the individual class period. The heart of their curriculum is a series of activities built around an explicit, step-wise modeling cycle that students complete with a particular problem (Lesh & Doerr, 2003). Their steps include (a) Description (mapping model from real world); (b) Manipulation of the model to make predictions about the real world; (c) Translation of those predictions to the real world; and (d) Verification of those predictions (and thus of the model itself) in the real world (p.18). This cycle has obvious ties to the many learning cycle approaches (such as the 5E model) for the teaching of inquiry in science, as students make the model (similar to Explain) from patterns observed in the real world (similar to Explore), then make and test predictions (similar to Elaborations) with their model, revising as necessary. These four steps are repeated over and over again. Six, one-hour modeling sessions are recommended for best results, with a different problem in each session (Lesh & Doerr, 2003).

In addition to the above curricular level approach for learning to model, Justi and Gilbert (2002a) also present a stepwise approach for teaching students to model at the individual lesson level. Their steps are (a) determine the purpose of the model, (b) observe the system and select the source of the data, (c) develop a mental model and revise as necessary, (d) test the model empirically, and (e) discuss scope and limitations of the model. The most unique step, as compared to the approaches in Lesh and Doerr (2003), is the final scope and limitations stage. Justi and Gilbert mention discussing scope and limitations as part of an advocacy stage to model

<table>
<tr><td></td><td>Trowbridge, Bybee, & Powell (2000)</td><td>Justi & Gilbert (2002a)</td><td>Schwarz & White (2005)</td><td>Lesh & Doerr (2003)</td><td>Kehle & Lester (2003)</td></tr>
<tr><td>Pre-modeling</td><td>Engage students in the problem, elicit prior knowledge</td><td>Determine the purpose of the model</td><td>Hypothesize</td><td></td><td>Simplification of a realistic problem</td></tr>
<tr><td rowspan="2">Creation of the model, typically through observation or other access to relevant data.</td><td rowspan="2">Explore the concept through observation and experimentation</td><td rowspan="2">Observe the system and select the source of the data</td><td>Investigate</td><td rowspan="3">Description (mapping model from real world)</td><td rowspan="2">Creation of a realistic model [mental]</td></tr>
<tr><td>Analyze</td></tr>
<tr><td>Explain what was learned, for instance, creation of a general rule</td><td rowspan="3">Develop a mental model and revise as necessary</td><td rowspan="2">Model</td><td>Abstraction to a mathematical model</td></tr>
<tr><td>Work with the newly created model, see what it can do.</td><td>Extend and elaborate. Explore near and far transfer of the knowledge, further development, different contexts</td><td rowspan="3">Manipulation of the model to make predictions about the real world</td><td rowspan="2">calculation of mathematical results</td></tr>
<tr><td rowspan="2">Test the model against reality</td><td rowspan="2"></td><td rowspan="2">Test the model empirically</td><td rowspan="3">Evaluate (includes, application to new situations, as well as scope and limitations)</td></tr>
<tr><td>Translation of those predictions to the real world</td><td rowspan="2">Interpretation of these mathematical results back in the context of the realistic problem</td></tr>
<tr><td>Post modeling</td><td>Evaluate</td><td>Scope and limitations</td><td>Generate new question</td><td>Verification of those predictions (and thus of the model itself) in the real world</td><td></td></tr>
</table>

Figure 2. Comparison of several approaches to teaching a modeling lesson.

good science practice, just as a scientist attempts to disseminate a model to the broader scientific community through publications and lectures. However, in modeling education this step could serve a further purpose, since this topic of scope and limitations of models relates to one of the most common difficulties students have with models, namely, the misuse of models in situations where the model is not suitable. Thus, by making this step an explicit part of a modeling curriculum, one could begin to address this misconception with students, and inculcate a habit of asking what is/are the scope/limitations of this model?

Kehle and Lester (2003) describe a similar modeling cycle that involves simplification of a realistic problem, creation of a realistic model, abstraction to a mathematical model, calculation of mathematical results, and interpretation of these mathematical results back in the context of the realistic problem (real world). This moving back and forth between the real and abstract worlds is strongly applicable to sciences such as chemistry, where chemists must move between the macroscopic real world phenomenon that is observed and either an invisible (particle) or symbolic world that can then explain the macroscopic phenomenon.

Schwarz and White (2005) make use of a learning cycle that closely follows scientific inquiry. Students first hypothesize about a question or situation. Students then perform experiments and analyze the data generated about the situation. Students construct a model from this analysis and evaluate this model in further investigation. Included in this evaluation was an application to other situations and a discussion of limitations. Students then use this model to generate new questions for investigations. This final step is the most significant, when comparing to other cycles, as one of the primary purposes of models in science is to generate new questions, and no other researchers mentioned in this review have taught this aspect.

Additionally, Schwarz and White (2005) predicted that engaging students in discussion and reflection would make students more metacognitive about modeling.

While not as comprehensive as the various learning cycles, other authors have specific strategies for addressing some aspects of modeling knowledge. Cartier, Rudolph, and Stewart (2001) find that students best understand models when working with models to predict and explain, and when refining models. Grosslight et al. (1991) specifically mention comparison of multiple models/representation of the same phenomenon as successful in increasing student understanding of models. Furthermore, by comparing two or more models, students are forced to accept that multiple models may exist and that models are not perfect copies, since if one model is better than the other, at least one of the two cannot be perfect (Lehrer & Schauble, 2003).

By constructing models themselves and making choices about what needs to be included in the model, students can come to understand that models are not exact copies of reality (Lehrer & Schauble, 2003).

All sources seem to agree, however, that modeling is truly iterative, requiring revision over many modeling cycles. Saari and Viiri (2003) add that if the gains in modeling knowledge are to be permanent, modeling needs to continue to be part of the curriculum, or the students will regress back to their previous modeling level.

Critical analysis of selected studies.

Three studies bear further scrutiny regarding the improving of modeling knowledge. Saari and Viiri (2003) present one of the most varied approaches to modeling. This research stands out from many others such as Valanides and Angeli (2006) or Ergazaki et al. (2005) in two important ways. First, while the others studied the use of a single modeling program

(*Model-it* or *ModelsCreator*), the students in the study of Saari and Viiri were exposed to a variety of different activities. Second, a pre- and post-assessment was used to determine gain in Saari and Viiri (2003), whereas the others (Ergazaki et al., 2005; Valanides & Angeli, 2006) only measured student success on the use of the program, with no pre-treatment measure.

The variety of the modeling activities used during the eight hours of intervention was strength of the curriculum. Kinesthetic modeling, black box activities, macroscopic and microscopic models of matter, and a computer simulation all presented students with different examples of modeling. The black box activity would help move students away from misconceptions such as models having a right answer (since the right answer was never revealed) and towards an idea of using indirect evidence. Movement between macroscopic and microscopic models and explanations of the states of matter should help students to think in terms the unseen theoretical agents described in Lawson et al. (2007). Limitations of models were discussed at each step, as well as individual written assignments forcing students to reflect on each of the models. These steps are in line with Justi and Gilbert (2002a) and the explicit-reflective approach of Akerson et al. (2000) respectively. Not surprisingly, positive results were achieved.

Using three levels similar, but not identical to, Grosslight et al (1991), 15 of 31 students moved one level and 14 students moved two levels. Even on the delayed posttest after three or seven months, a net gain of 30 levels for 31 students was obtained. While these results are encouraging, there are several limitations to this study. First, the sample sizes are quite small, consisting of 14 and 17 students. Second, Figure 6 on page 1344 does not match the text, as it appears School A and School B are flipped. Since these schools showed marked differences in performance, and since one of the author's was the instructor for this research, this is not a trivial

mistake. Differences in performance between the two groups is explained as having to do with later science instruction, but implementation validity or other explanation relating to the researcher/teacher could be as likely.

The next study also shares the researcher/teacher aspect, but the methodology and clarity of presentation help to minimize its effect. Windschitl and Thompson (2006) studied 21 pre-service teachers. Like Saari and Viiri (2003) these students were engaged in a variety of learning activities, including computer simulations and labs where the mathematical rule generated was referred to as a model. Students also conducted an in research project that was intended to be model-based, and to present on it at the end of the course.

The mixed-methods data collection was impressive as eight data sources were collected and analyzed. An initial questionnaire and end of course questionnaire bracketed the instruction, and these were coded on a three-point scale similar to the one used in Grosslight et al. (1991). A number of qualitative sources were also used including journals, reflections on activities, videotapes of inquiry investigation presentation, transcripts of class discussions, the student's modeling lesson plans, questionnaires about previous inquiry experience, and even records of informal conversations during the six month course.

Another strength of this study was a rubric for the inquiry investigation and presentation that required students to show how the model was used in the reasoning. Many students carried out what they felt were successful investigations based on their previous conception of science. These students were able to establish a relationship between variables, but were not able to use model-based reasoning and evidence from their investigation to support their model, or their model to generate testable hypotheses.

The quantitative gains seen were not as large as the gains in Saari and Viiri (2003), with only six levels moved (net) over 21 students, or just a little less than a third of a level per student. However, this study looked at a sample where over 1/3 of the students were already at the maximum score possible, whereas in Saari and Viiri (2003), 29 of 31 were at the minimum possible score, so gains were easier. Calculation of normalized gains would have made any comparisons between such studies more equitable. The fact that four students' knowledge of models went down over the course of the class is also a concern. Especially when students' modeling knowledge appeared to have dramatically increased after the technology portion of the class, what was the nature of these lower scores? Did they represent new misconceptions, a lack of effort on the posttest, or something else? The authors contend less thorough answers on the posttest. Ceiling effects too could have played a role, for those at the highest level had no place to go but down. Since no mention of methods used to ensure pretests and posttests were scored with the same rigor, it is also possible that stricter standards were used on the posttest, so as not to commit the bigger bias error of showing a gain where there is none.

While not a weakness of the study per se, given the low success rate (two of 21 students completed a model-based inquiry), if these ideas were to be applied to a classroom setting or similar research was to be attempted, the instructor should carefully consider limiting topics to those that could at least potentially lead to model-based inquiry.

One interesting finding from Windschitl and Thompson (2006) is that students can talk sophisticatedly about models, and not have any idea how to use them in a scientific investigation. Another interesting relationship was that students who had a very strong background in the scientific method as taught traditionally in schools were among the most resistant to learning how to perform model based inquiry.

The third of the three model based studies also involves a teacher researcher in a pre-service teacher method's class. Like Windschitl and Thompson (2006), Cullin (2004) collects a large variety of information, including process video of students building models, the models actually generated, video of classroom activities, student artifacts, pre- and post-modeling questionnaires and pre- and post-modeling interviews.

One limitation of this study is the severe limitation on time for what may be a fairly new task for the learner. Students must build the models within the confines of the classroom, over the course of two, one-hour sessions, in front of a video camera, with a partner assigned by the researcher, using a piece of unfamiliar software, *Model-it*. How important is this time constraint on a student's opportunity to learn from the modeling experience?

One strength of this study, compared to Windschitl and Thompson (2006) and Valanides and Angeli (2006) is that the students all were required to build the same model, rather than having the freedom to select their own topics. Moreover, they were all given the same concrete experience with the phenomenon they were modeling (a pond). This uniformity of task made for many fewer extraneous variables that could influence the quality of models. Because there was a standard model and a standard rubric for scoring this model, students were not limited in creation of their model by their chosen topic. Because the rubric used rewarded the number of variables and relationships created within a model, it would only be appropriate in situations where students were modeling the same or similarly complex phenomena. A disadvantage to having all students create the same model is that students would be more likely to collaborate across the class on their projects. While videotaping would minimize or at least alert the researcher to the extent that collaboration between pairs was occurring inside of class, the fact that this modeling project spanned two sessions would not prevent it from happening outside of class. The pairs

themselves also provide an advantages and disadvantages in that they encourage students to externalize mental processes by communicating with their partner, yet may mask the true extent of each partner to model on their own.

Relating the Variables

While no studies that attempt to link all three variables (level of cognitive development, knowledge of models and understanding of the nature of science) appear in the literature, there are a number of studies that attempt to link two of these variables. Asami, King, and Monk (2000); Lesh and Doerr (2003); and Lesh, Cramer, Doerr, Post, & Zawojewski (2003) will be used to examine the relationship between cognitive development and models. Lawson's studies relating some aspects of the nature of science to cognitive development will be discussed. A few final points on the relationship between modeling knowledge and the nature of science not presented elsewhere are examined at the end of this section.

Cognitive development and models.

A number of studies have presented a relationship between modeling ability and development, although the relationship was not the focus of the research; the exception is Asami et al. (2000). In other studies the most common relationship appears to be that students have a better conception of models at an older age (Grosslight et al., 1991). Since models tend not to be taught explicitly in many curricula, this growth in modeling knowledge could reflect a change in cognitive structure and/or epistemology. Two different historical perspectives take opposite views on this relationship. The Vygotskian tradition tends to see the external models as representations of existing internal structures, whereas the Piagetian tradition tends to see the conflict of working with models as leading cognitive development. Either way, "Cognitive

development is closely related to the ability to represent … either internally or externally and to move successfully between the two," (Sakonidis, 1994, p. 39)

Bliss (1994) commented on the thoughts of Vygotsky on modeling. Vygotsky sees a relationship between models and cognitive development, saying "The process of internalization [of a model] is not the transferal of an external activity to a pre-existing internal 'plane of consciousness': it is the process in which this plane is formed" (Leon'ev, 1981, p. 51 as cited in Bliss, 1994, p. 29). In addition, Vygotsky feels that both speech and modeling are externalized thought. "Good modeling tools will present learners with structures that helpfully allow their thoughts to find expression" (Bliss, 1994, p. 31). Another challenge is knowing what variables from the dense real world to incorporate into a simplified model; and likewise, what to leave out. Testing a model is cognitively demanding and really should not occur before age 15, as it requires higher order thinking skills such as separation of variables. To a pupil, "a model looks not like a thought but like a thing" (Bliss, 1994, p.32). This quote implies why students may have difficulty appreciating non-physical models. However, Vygotsky (1981) also felt that externally mediated representations preceded internal development. Others have had similar comments regarding modeling and social constructivism. The internalization of the external dialogue in the social construction of models leads development (Lesh, Cramer et al. 2003). Students can monitor the behavior of others (in a group) before they can modify their own (Lesh, Cramer et al. 2003), thus critiquing the models of others might be a stepping stone to creating one's own model.

Much of Piaget's work regarding models has already been addressed. Campbell and Olsen (1990), in the Piaget tradition, state that creating models and other external manifestations indicates pre-existing internal cognitive structures. As discussed previously in the modeling

section, the natural progression is from qualitative to quantitative modeling. This progression follows Piaget's cognitive levels. In the height-mass example, a student looking at the raw data will first notice the qualitative pattern that both columns increase. Next, a student will notice an additive pattern. The numbers in the Height column increase by three cm every time, and the numbers in the mass column increase by a little less than one in most cases. Additive reasoning is indicative of the pre-operational cognitive level. However, further investigation reveals that an additive pattern is insufficient to explain the Mass column, as the student may notice that the gap between consecutive masses increases a little each time, from 0.6 kg between the first two entries, to 1.1 kg between the last two entries. Thus, the student arrives at a semi-quantitative conclusion that as height increases, mass increases faster and faster. Finally, a regression reveals the fully quantitative relationship, typical of formal operation reasoning, given by the equation $\text{mass} = 0.0061*\text{height}^{1.707}$. This progression in reasoning from the qualitative to the quantitative is found multiple times by Lesh and Doerr (2003) in their observations of students completing modeling activities.

As mentioned earlier with regards to qualitative, quantitative, and semi-quantitative modeling, Lesh and Doerr (2003) found in their modeling curriculum that during the course of modeling a problem, students typically went through the same stages each time. These stages were qualitative reasoning, using only a subset of information, additive reasoning, sometimes primitive multiplicative reasoning, and finally pattern recognition. They found that full multiplicative proportional reasoning (defined as a second-order relationship, or a relationship between relationships) may or may not be met in every case. However, Lesh and Doerr (2003) find a well-defined relationship between the local conceptual development of their students and the general cognitive development as described by Piaget. In other words, within each modeling

activity, students progress from the concrete to the formal. However, as each new problem is presented, students progress once again (although not necessarily at the same pace) through each stage of development (Lesh & Doerr, 2003). Even students who had reached the formal level in a previous activity start at the concrete level in the next activity. Lesh and Doerr noticed that students may regress to a lower stage as features of the problem change, even if the features are not central to the problem, which is similar to the expert/novice dynamic mentioned previously in which students identify problems by surface similarities instead of conceptual similarities (Chi et al., 1981). Metacognition, which may be a necessary component to enable students to make the modeling process more accessible and useful, appears not to occur unless the investigators provide specifically for it in the activities (similar to the reflective explicit method discussed previously for teaching NOS). The model development activity sequence mirrors much of Piaget's perspective on cognitive development (Lesh & Doerr, 2003). The most striking difference is that Piaget sees general conceptual development organized in ladder-like stages, while Lesh and Carmona (2003) see modeling in highly specialized conceptual systems, and local development within each problem solving session.

While modeling seems to be tied to higher levels of cognitive development and older students in theory, in practice, modeling activities have been attempted with younger students with some success. Age and/or cognitive development does not appear to be the only factor determining the success of modeling, as successes and failures have been found across a variety of samples. However, the reason for the failures is often not addressed, and this could be related to developmental level.

If science and math are really about modeling reality, then instruction in modeling should begin before high school (Lehrer & Schauble, 2003), for even young children can abstract that a

stick is a sword or a banana is a telephone. Ratios, which are a relationship between two variables and thus a stepping stone towards mathematical modeling, can be introduced as early as third grade (Lehrer & Schauble, 2003). Thus important modeling precursor skills are being built well before students are able to reason formally. Bliss (1994) takes this idea one step further. They claim Piaget undervalues the concrete thinking as only a step towards formal thinking and contrast Piaget's ideas with those of Johnson-Laird (1983), who sees concrete models as useful in and of themselves. How the model is used appears to be the key, for it is the relational structures and analogical reasoning that can be built upon, rather than the concreteness of the model, which should be the focus. This theory is put in practice in the next study.

Ergazaki, Komis, and Zogza, (2005) find in their study of 36 12-year-olds that these students were able to construct models with some success. These students are able to create models. Fourteen of the 18 pairs of students are able to complete a model, with 12 of the 14 pairs able to link at least five variables, and five pairs of students are able to link between eight and 12 variables. However, the models appear to be limited by the students' low-level, convergent thinking. The authors explain that convergent thinking (which is likened to brainstorming) results in lists of variables that are related, but does not provide further organization.

This same study (Ergazaki et al., 2005) refutes the idea that young students are capable of true formal modeling when the authors investigate the quality of the models. Ergazaki et al. (2005) investigate the level (macroscopic, microscopic) of the variables in the student-created model. Most of these young students do not link microscopic variables with macroscopic variables. Instead, they tend to link macroscopic to macroscopic and microscopic to microscopic, with the only exceptions being concepts like water (H_2O) that can be perceived of

as a macroscopic substance, but also can appear in the chemical equation (microscopic) for photosynthesis. However, a microscopic (and somewhat abstract to students) concept like photosynthesis is not linked to a macroscopic concept like plant growth, a serious conceptual flaw. Furthermore, as many of these variables are provided to students within the software, these students are not building a model from scratch, as detailed previously in the section on about ways to teach modeling. The brief nature of the modeling activity, one hour of instruction and practice followed by one hour of modeling, further limited the conclusions that can be drawn from this study.

From a more practical standpoint, relevant experience may prove to be a limiting factor to a student's ability to model. One important step in modeling is deciding what factors to include in the model. However, what goes into the model depends as much on previous knowledge as it does on the nature of the system being represented. "We make what we can model, not model what we fancy making" (Ogborn & Mellar, 1994, p.19). Therefore, the relevant life experiences gained with age may have as much to do with older students' successes at modeling as the progression of cognitive development that is supposed to come with age.

Asami et al. (2000) study the relationship between students' level of cognitive development and the students' mental models of electrical circuits, although this study does not work extensively with different representations of these mental models. This quasi-experimental study involves an experimental and control group of 10-11 year old students in a Japanese government school in London, England. Students in both groups act out circuits carrying electricity and complete the same seven lessons from the standard Japanese curriculum. Unlike the control group, the experimental class of students is given direct instruction in the exact circuits that will be used on the delayed posttest. Therefore, it is expected that students from the

experimental group will perform better on the delayed posttest, since they have already seen the questions and answers, and assuming memory of the instruction will play the dominant role in the students' answers. In fact, the experimental group did not perform better. One question is used as a control for this recall, and requires students to trace the path of electricity like the circuit both groups acted out; both experimental and control groups responded similarly. However, the other questions on the test were seen and discussed by the experimental group, but not the control. Answers to four of these five questions were also not significantly different, with $p > .2$. Only one question yielded significantly different answers, and it is hypothesized that this is because the experimental group's experience moved them away from the mental model represented by the most frequent wrong answer, and towards a more scientific model, although it is not clear why. Perhaps the mental models in question are relatively similar developmentally, and the direct tuition is able to achieve some motion in their mental models whereas the other mental models are beyond these students cognitive developmental level to understand, direct instruction or not. Furthermore, when the questions and answers are used to determine the dominant mental model of the experimental students, the percentages of students holding each mental model correspond well to the results of measuring cognitive development by Shayer and Adey (1981), implying a certain cognitive level may be necessary to use a particular mental model for electricity. The fact that the experimental group appears to benefit in one case from tuition is attributed to local cognitive development similar to what has been reported by Lesh and Doerr (2003).

While the Asami et al. (2000) study seems to be the necessary impetus to further research in this area, no additional studies appear to follow. Unfortunately, this single study (Asami et al., 2000) with small sample size and limited context is the only study available linking the

understanding of models to cognitive development. The authors themselves lament this fact, stating "researchers have been content to document the variety and occurrence of different mental models rather than consider the reasons for those specific proportions [of students using each model]. Here, following Monk (1990, 1995) we suggest that an account can be given and that such an account needs to draw on ideas of cognitive processing" (p. 151).

Many other studies show relatively high failure rates with models, yet fail to attribute these rates to a lack of cognitive development. Ergazaki et al. (2005) shows 22% of their 12-year-olds cannot model at all, with another 22% modeling only at a macroscopic (which would seem to corresponds to concrete) level. Valanides and Angeli (2006) show 28% of their pre-service teachers are unable to identify appropriate variables for their model correctly, while only 13% construct models that are correct in structure and relatively complex. Windschitl and Thompson (2006) also report that only two of 21 students in their study construct true scientific models capable of generating testable questions, with seven more at least proceeding to a stage where relationships are determined from empirical evidence and the resulting models are tested for accuracy. Almost half (10 out of 21) do not construct models that can be empirically tested.

While other studies explicitly linking models and cognitive development do not exist, as has been pointed out previously, practicing scientists may see models and theories as interchangeable. Thus, studies relating cognitive development and the nature of science have significant bearing on the relationship between models and cognitive development as well. Thankfully, these studies are more numerous, especially when one considers the nature of hypothesis formation and testing as central to the nature of science and scientific models.

Cognitive development and nature of science.

Anton Lawson's work in this area is important. In Lawson, Clark, Cramer-Meldrum, Falconer, Sequist & Kwon (2000), the relationship between students' ability to reason and test hypotheses and their cognitive development is investigated. One of the classic tests of cognitive development in Piaget and Inhelder (1955, 1966) requires students to identify variables that influence the period of a pendulum. In this activity, the students are able to directly change (or see a change in) the independent variable, while at the same time seeing a change in the dependent variable. Students at the formal level are successful at this task, and can make correct hypotheses and interpret these hypotheses in light of experimental evidence. However, Lawson et al. (Lawson, Clark, Cramer-Meldrum et al. (2000) find that many students who are successful on problems similar to the pendulum problem have a very low success rate on problems in which a choice of mechanisms involving *unseen* agents is involved. Further investigations reveal that the ability to reason formally, and more importantly, what Lawson, Clark, Cramer-Meldrum et al. (2000) call post-formal reasoning, is a better predictor of students' ability to answer questions designed around specific cognitive abilities than the amount of declarative knowledge the students possess. To use a computer analogy, it is a question of whether the student had a 64-bit or a mere 32-bit processor that is more indicative of success on these problems (and in class) than the amount of data stored on the hard drive.

Nature of science and models.

The relationship between the nature of science and modeling is strong but limited almost exclusively to that part of the nature of science relating to theories and their revision. Take for example, the quotes "modeling is a central skill in scientific reasoning," (Forbus, K., Carney, K., Sherin, B., & Ureel, L, 2004, p. 1) or "models are both the methods and the products of science"

(Harrison, 1998, p. 420). Other aspects of NOS, however, are at least tangentially related to modeling. These will be addressed first, since they are rarer.

Through choices in the relational structure of data, a modeler is able to make answers emerge and disappear. Through asking the right questions and gathering and including an essential variable in a model, a modeler can bring forth answers that were hidden in models that did not include this variable (Lehrer & Romberg, 1996). This effect could be seen as a manifestation of creativity in the nature of science (creativity being one of the major categories tested the SUSSI created by Liang et al. (2006)).

However, most of the relationship between NOS and models revolves around theory building. "It [modeling] tries to give children as much freedom as possible to manipulate those ideas, both in order to help them understand the world better, but also in order to lead them to an understanding of the nature of the task of theory building itself," (Mellar & Bliss, 1994, p.1). Thus, modeling serves multiple purposes, according to Mellar. It serves content purposes by increasing students' understanding of the world by manipulating ideas, but it also serves nature of science goals by engaging students in more authentic science tasks. Notice that Mellar starts by talking of models, but ends by talking of theories; this demonstrates the virtual equivalence of these words to him. Wisnudel-Spitulnik, Kracjik, and Soloway (1999) observe strong NOS growth in some students after completing modeling activities related to the environment, particularly with respect to the roles models play in generating testable hypotheses. In addition to the direct benefits of improving student understanding of models, a modeling curriculum is also mentioned as strengthening scientific inquiry (Cartier et al., 2001). The authors (Cartier et al., 2001) appear to equate scientific inquiry with something similar to the student process skills using the scientific method.

Akerson, Abd-El-Khalick & Lederman (2000) reveal some relationships between the nature of science and the nature of models in their study. Several explicit connections are made between NOS and models when discussing the inferential nature of science (atoms cannot be studied directly, and models of their structure are built on inferences made from indirect observations) as well as two activities, the tube and the cube, in which students must make assumptions about an unknown part of the apparatus from observing the visible parts and certain behaviors. In the tube, students construct a model of the insides of the tube that appears to behave the same as the real tube, but cannot know if their model is an exact representation or not, nor does it matter, as long as it functions correctly.

Conversely, it may be that students' difficulties with scientific models have as much to do with a misunderstanding of NOS as a misunderstanding of models. Windschitl and Thompson found that strong adherence to a school science approach limited students ability to learn from a model base inquiry. Student confusion about the nature and goals of science leads to student difficulty with science (Reif & Larkin, 1991). Reif & Larkin's arguments are summarized in Table 1.

Many of the authentic science ideas such as iterative process, propositional knowledge, connection of facts, and application of ideas to new data sets are, not surprisingly, very much in line with scientists' views of models. Likewise, everyday and school views on science such as absolute truth and discrete facts are so incompatible with modeling as to limit models to merely very faithful physical models.

Furthermore, school science is neither real science (as practiced by scientists) nor everyday experience, and falls somewhere between the two, adding another obstacle to student understanding (Reif & Larkin, 1991). Teaching students about the process of science is pointless

Table 1. *Comparison of science and school science/everyday experience.*

	Science	School science and/or everyday experience
Purpose	Science is about connecting facts and theory building	Simple amassing of facts
Role of models	Thinking tools in science used to make predictions	Visualization
Role of truth	Accept approximations of the truth that work well enough	Absolute truth
Number of rules	Parsimony, i.e. to make a maximum number of correct inferences from a minimum number of rules	As many rules as are needed
Length of inference chain	Greater distance between the rule and the phenomenon it explains because of parsimony	Relatively short
Type of knowledge	Propositional (relationship between facts) and procedural (such as how theories and models are refined)	Memorization of factual knowledge
Coherence	Scientific knowledge must be coherent	Rules may change
Source of truth	Observation is the ultimate arbiter of the validity of a rule	Many sources of truth

Table 1. *Continued.*

	Science	School science and/or everyday experience
How knowledge is acquired	Iterative process of successive approximations and refinements Scientists are continuously attempting to apply their theories or models to broader data sets, and modifying the theory as appropriate	Linear process

discussing the predictive nature of models, or approaching level three modeling knowledge on the scale used by Grosslight et al. (1991). The analysis of the *Purpose or Utility of Models* sub-score, similar results were found. Two of the five questions showed significant gains with three not showing gains. Again, ceiling effects were given as the reason for two of the three areas not showing gain.

Both the *Nature or Process of Modeling* sub-score and the *Evaluation of Models* sub-score showed Cronbach's alphas $\leq .20$, so no further statistical analysis was done at a sub-score level. Both sections showed questions where students scored lower on the posttest than on the pretest, specifically on questions related to models omitting aspects of the phenomenon that are not necessary and the meaning of multiple models for the same phenomenon.

Therefore, the significant gain in modeling knowledge claimed demonstrated by this instrument in this study seems to in large part rest upon a very large gain shown on one question,

without reinforcing this process by repeatedly practicing it (Reif & Larkin, 1991). Modeling, with its iterative nature, is a good tool for reinforcing the nature of science.

Analysis of Schwarz and White (2005).

Schwarz and White (2005) is the final article that will be critically analyzed. It claims gain in scientific inquiry knowledge, physics content knowledge, and modeling knowledge using an extended (10.5 week) modeling curriculum. It uses pretest and posttests to establish modeling gains, and attempts to triangulate these with clinical interviews regarding modeling. An inquiry assessment was used (again, pretest and posttest) to support claims regarding scientific inquiry. Scientific inquiry is related to, but not identical to, the nature of science. A physics knowledge pretest and posttest was also conducted.

Significant gains in all three areas (modeling, inquiry, and physics content) were reported. Correlations between the three posttests were all significant with $p \leq .01$. Thus, a modeling approach can be used to improve all three types of knowledge and gains in each are related. A closer look at some of the gains, however, paints a less clear picture.

Particularly with regards to modeling knowledge, the significance of the gains is dubious. The total modeling gain (from a mean of 61% on the pretest to 70% on the posttest) was significant ($p \leq .001$). However, only two of the four sub-scores of the test (nature of models and purpose of models) showed gains. Examining further, in the sub-score concerning nature of models, huge gains (30% or more) in the questions regarding understanding types of models overwhelmed the fact that five of the six other questions making up this sub-score showed no gains (including questions regarding multiple models and the constructed nature of models). This lack of gain was attributed to ceiling effects, but perhaps represents the difficulty of teaching these more conceptual aspects of models than the types of model question that showed

such gain. In fact, the only other question showing significant gain in this sub-score was the question regarding a definition of model, again, a fairly low level concept. The interviews did support that student definitions of models were in line with scientists' definitions, with 64% and significant gain on three others, with 14 other questions showing no significant gain, and in some cases, a loss. However, as is pointed out, perhaps a better instrument without these ceiling effects would have been capable of showing greater gains across all questions and sub-scores.

While not addressed as such, there is an indication from one small part of the stated results that cognitive development may have played a role in the success of their curriculum. Schwarz and White report that students scoring below 60th percentile on the Individual Test of Academic Skills (a variable used in their statistical analysis) only showed gains of 3% in modeling knowledge compared to students who scored above the 60th percentile, who showed a gain of 11% in modeling knowledge from pre to posttest. Could better cognitive development of the high achieving students explain why they benefited preferentially from intervention?

Conclusion

In conclusion, the review of the literature reveals that there are existing instruments and procedures to measure student knowledge of the nature of science, student conceptions of the nature of science, and students' cognitive developmental level. The variety of these instruments allows instruments to be selected that meet the needs of the study while staying within the appropriate time constraints for a class which has neither nature of science nor modeling explicitly as a goal.

Methods for improving student understanding of the nature of science, student conceptions of the nature of science, and students' cognitive developmental level were also revealed. The explicit reflective technique appeared to be more successful in teaching the nature

of science than approaches requiring students to abstract the nature of science from inquiry practices. Methods for teaching modeling center on step-wise or scaffolding techniques, as well as learning-cycle approaches. While not measured in this study, long term gain in cognitive developmental level has been achieved through practice working through more cognitively demanding problems. Thus, the literature seems to imply a gradual approach to modeling, with explicit reflection on what was learned about models and modeling at each step.

The literature also reveals that there are strong relationships between these three variables. The links between models and theories, and model building and theory building are well established. Links between cognitive development stages and the steps to constructing a model appear to be more than coincidence, but appear to raise the question of local versus global cognitive development. That many modeling studies show a sizeable failure rate, and that in one case, failure to adopt and use particular models appears linked to cognitive development completes the chain of logic.

CHAPTER THREE

METHODOLOGY

As discussed in the literature review, a considerable number of studies exist at a variety of levels which have attempted to describe a person's understanding of modeling and ability to construct models. The nature of science is another well-developed area of research, and various approaches have been developed to measure understanding of this concept. The area of Piagetian development also has well-developed instruments and protocols. This study attempted to establish a relationship between the three areas, a study demonstrated to be lacking from the literature. Does a student's Piagetian developmental level influence the extent to which a curriculum designed around several incrementally more complex modeling activities results in deeper understanding of models and the nature of science, specifically: (a) the relationship between theories, laws, models, and hypothesis; (b) how and why theories change over time; (c) how and why models are refined; (d) the purposive nature of model creation; and (e) the role of models in scientific investigations?

Research Questions

Is attainment of the formal operational Piagetian level of understanding necessary for a model-based environmental science curriculum to increase students' understanding of models and the nature of science?

Sub-questions.

1. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations), and is that improvement related to Piagetian level?
2. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the utility of models (communication, simplification for study, prediction), and is that improvement related to Piagetian level?
3. Does a curriculum emphasizing student comparison, refinement, and creation of models improve student understanding of the relationship between models, theories, and the scientific method (models operationalize theories, allowing them to be tested with the scientific method), and is that improvement related to Piagetian level?

Hypotheses

Null hypotheses.

There will be no significant or important difference at the $p = .05$ level in student understanding of models nor understanding of the nature of science before and after completing a semester of the model-laden environmental science curriculum. There will be no significant difference at the $p = .05$ level between any normalized gain between the pretest and posttest in modeling and/or nature of science knowledge between students in the post-formal operational stage, formal operational stage, early operational stage, and pre-operational/concrete stage of cognitive development. (This curriculum included exposure to authentic model use, critique and modification of existing models, comparison of multiple models of the same system, analysis of

the conscious choices that shape models, and construction of models and use of these models to answer questions.)

Alternative hypotheses.

There will be statistically significant gain in students' modeling knowledge and/or nature of science scores on the posttest as compared to the pretest. This difference will also be statistically important, showing a normalized gain of greater than 0.5 (medium effect). Furthermore, when any gains in modeling and/or nature of science knowledge are correlated to the cognitive development of the same student, it is expected that students who have reached a higher operational level of development (post-formal > formal > transitional > pre-formal) will have statistically greater gains than students with lower levels of development.

Methodology

The methodology drew upon accepted mixed-methods approaches for gauging the students' knowledge of the nature of science and modeling both prior to and after taking Chemistry 304 (see The Setting, following, for more detail). Cresswell (2003) states four considerations for mixed-method research: implementation, priority, integration, and theoretical perspective. The implementation strategy was for sequential data collection, with quantitative pretest and posttests bracketing primarily qualitative methods throughout the course. Priority of the quantitative and qualitative portions of the study was considered equal initially, with qualitative methods providing clarification of and explanation for any trends, expected or unexpected, in the quantitative data. In practice, the data and analysis presented in chapters four and five are almost exclusively quantitative, with only select qualitative pieces used. Information from both quantitative and qualitative sources was integrated at a number of levels. The SUSSI and SUMS tests used to measure the dependent variables have both Likert-scale and

free-response questions and so represent integration at the data collection level, with an attempt at triangulation between each instrument. Integration also occurred at the data analysis stage, as qualitative information gathered from individuals in class activities and reflections, as well as in the final modeling project, was used to triangulate the results of the quantitative posttest. As the validity of the instruments used to measure the dependent variables is in questions initially, establishing this triangulation of the SUMS to the dominant measure of modeling ability in the literature, the three levels of Grosslight, Unger, Jay, and Smith (1991), is essential.

Theoretical perspectives.

The initial theoretical perspectives follow deductively from the ideas that have shaped the literature review, repeated here. The first theoretical perspective links cognitive developmental level and modeling; the second, links modeling and the nature of science. Cognitive development of human beings appears to develop in stages, first established by Piaget. Because students develop at different rates based on experiences and other factors, samples of students, especially in intact classrooms, should represent a cross-section of cognitive abilities. Attainment of the formal operational level of cognitive development is necessary to reason abstractly. Models, particularly scientific models, are abstractions. Nearly every modeling study revealed a number of students who were unable to model. Developmental level may be the variable that explains the heretofore unexplained but ubiquitous failure to model of a fraction of students within each study. Since models and theories are deeply interrelated in science, growth in modeling ability and knowledge should be linked to a better understanding of theories, an important nature of science component.

The Setting

Chemistry 304: The Environment and You. At the institution where this study took place, all students are required to take one class addressing Learning Area 10: People and the Environment. Chemistry 304 is an Outer Cluster class, meaning it is designed to be taken after taking a minimum of one class each in oral communication, written communication, mathematics, critical and multicultural thinking, and a natural science (MSUM, 2006). In Chemistry 304, concepts of man's relationship to the environment from a chemistry perspective are explored. Chemistry 304 also has a writing intensive designation, meaning at least 16 pages of formal writing in multiple drafts must be completed.

Pilot

A pilot study was conducted during Fall Semester, 2007, to examine the overall feasibility of the design. The instruments were carefully tested and examined with a sample of 25 students in a Chemistry 304 classroom, including an analysis of question wording, sequence, scoring rubrics and exemplars (for additional details, see Appendix B: Achieving Inter-rater Reliability). The instructional approaches and student activities were also tested and refined for suitability.

Research Design

This study, like most modeling studies discussed in the literature review, is a mixed method study, but differs from most in that it takes a stronger quantitative stance. Primarily, quantitative data from Likert-scale and multiple-choice questions and written responses quantified by a rubric constitute the bulk of the data that was analyzed. Specifically, these consist of the Student Understanding of Science and Scientific Inquiry (SUSSI) Questionnaire (pretest and posttest) (Liang, Chen, Chen, Kaya, Adams, Macklin, & Ebenezer, 2006), Students'

Understanding of Models in Science (SUMS) (Treagust, Chittleborough, & Mamiala, 2002) (pretest and posttest), and Lawson Classroom Test of Scientific Reasoning (CTSR) (pretest only). It was assumed for this study that cognitive level would remain constant; however, and this assumption was to be tested with a follow up posttest of a subset of students. This check was not completed due to a lack of available volunteers after the course was completed. While Lawson, Alkoury, Benford et al. (2000) have found statistically significant gain in student's scientific reasoning as measured by the CTSR (effect size = .87) during a curriculum specifically designed to increase scientific thinking among college biology students, Adey and Shayer (1990) critiqued Lawson's previous work (Lawson & Snitgen, 1982) for not demonstrating "transfer to schemata not included in the program" (Adey & Shayer, 1990, p. 268). Likewise, Adey and Shayer (1990) categorized much of the research on gains in developmental level to that point as training rather than learning, reflecting the lack of general transfer of these cognitive abilities to other tasks. Their studies (Adey & Shayer, 1990; Shayer and Adey, 1992a, 1992b, 1993) showed that an intervention involving 30 activities and spanning two years was capable of increasing cognitive development across a variety of disciplines. Therefore, the conservative stance would be that this course of a mere semester would not significantly increase student's cognitive development during the brief nature of the intervention, and is the default stance since this was not measured with a posttest. The scores on the modified SUSSI and SUMS provide the primary dependent variables for this study, and it is hypothesized that there would be change in these scores. A "one-group pretest-posttest design" (Cresswell, 2003, p. 169) was used to measure gains in modeling ability with the modified SUMS questionnaire and nature of science knowledge with the modified SUSSI questionnaire. With respect to the more qualitative information, this study can be considered a case study design (Krathwohl, 1998) in that only

posttest measures are used, where the case is the Chemistry 304 classes during Summer Session, 2008 and Fall Semester, 2008.

Instruction

On the first day of class, a very general review of science as a way of knowing, the scientific method and experimental and control variables, were presented. The way in which experimental methods may differ in environmental science from other sciences students may have taken because there is not a *control* earth and an *experimental* earth for studies such as global warming, and the fact that it is not always ethical or desirable to run control studies on, for instance, pollution's effect on humans, was also discussed. Other than this initial instruction, there was no direct instruction in the nature of science. Since the pretest came after this instruction, any gain in the nature of science shown by the SUSSI scores may be attributable to the modeling activities. The timelines for the Summer and Fall Semesters are presented in Table 2 and Table 3.

Pretesting. The SUSSI and SUMS pretests were administered during the first class period in a computerized format. Each of the questions was entered into the "quiz" feature in Desire 2 Learn, with the Likert-type questions entered as multiple choice questions where the most scientifically accepted answer receives one point and each successively less acceptable answer receives 0.25 points less to a minimum of 0 points for the least acceptable answer. The free response questions were created as free response questions in D2L with an unlimited text box. These questions were hand scored according to the rubric. The computerized format was chosen to expedite data analysis.

During the second class period, students were given a printed version of the Lawson Classroom Test of Scientific Reasoning with a Scantron answer sheet and as much time as they

Table 2. *Study Timeline for Summer Semester, 2008 (Each day is a 110 minute class)*

Course Day	Major Task	Modeling component	Data Collected
1	Introduction to class		SUSSI Pretest, SUMS Pretest, Informed Consent
2	Introduce 2 non-physical models.	Tragedy of commons simulation (model); matter cycles (C, O, N, P)	CTSR
3			Start Follow up interviews.
4	World population data sheets activity	Models of population growth, factors that effect it, different assumptions lead to different predictions	
5			Reflections of world population data sheets
6	Food and water use activity	Use and critique of a mathematical model	
7			Critique and reflections of food and water use activity

Table 2. *Continued.*

Course Day	Major Task	Modeling component	Data Collected
9	Carbon Footprint modeling activity Day 1.	Use, comparison and contrast, critique of multiple mathematical models of same target.	
10	Carbon Footprint Modeling activity, Day 2.	Analysis of the construction of at least 1 of these models. Practice constructing own with spreadsheet software.	Comparison, reflections, critique of carbon footprint model
12	Global Warming Model	Analysis of a probabilistic model to make predictions. Analysis of underlying assumptions.	Reflection on global warming models
14			Topic of model due
16			Concept map of model due
17			Draft of model due.
18			Final modeling project due.
19			SUSSI and SUMS Posttest
20	Final Exam		Start post interviews

Table 3. *Study Timeline for Fall Semester, 2008 (each day is a 50 minute class)*

Course Day	Major Task	Modeling component	Data Collected
1	Introduction to class		SUSSI Pretest, SUMS pretest, Informed Consent
2-3	Introduce 2 non-physical models.	Tragedy of commons simulation (model); matter cycles (C, O, N, P)	CTSR
6			Start Follow up interviews.
8-9	World population data sheets activity	Models of population growth, factors that effect it, different assumptions lead to different predictions	
10			Reflections of world population data sheets
14	Food and water use activity	Use and critique of a mathematical model	
15			Critique and reflections of food and water use activity

Table 3. *Continued.*

Course Day	Major Task	Modeling component	Data Collected
20	Carbon Footprint modeling activity Day 1.	Use, comparison and contrast, critique of multiple mathematical models of same target.	
21	Carbon Footprint Modeling activity, Day 2.	Analysis of the construction of at least 1 of these models. Practice constructing own with spreadsheet software (planned, but time ran out).	Comparison, reflections, critique of carbon footprint model
24			Topic and concept map of model due
26	Global Warming Model	Analysis of a probabilistic model to make predictions. Analysis of underlying assumptions.	Reflection on global warming models
33			Draft of model due.
36			Final modeling project due.
37			SUSSI and SUMS Posttest
40	Final Exam		Start post interviews

need to finish the test. During the course, students were exposed to a variety of models and modeling activities (Appendix D contains the actual student directions for all major modeling activities). The Tragedy of the Commons was a simple physical simulation where students represent users of a renewable natural resource, such as fish, which were represented by pieces of paper on the table. A candy bar represented a paycheck that the student could get if they have ten fish to trade. Students could take as many or few fish as they wanted to, each turn, and at the end of the turn, each fish that is left reproduced and the fish count doubles. The first time the activity was done, since communication was barred, typically at least one person made a grab for the fish at some point to complete his 10, and the fishery died. In further runs, cooperation was allowed and students set fishing quotas.

An early activity (Human Population Lab) was the student's first exposure to mathematical models in this class (The Human Development Index and the Gini coefficient, a measure of income inequality). The specific numbers yielded by these models tended to elicit a desire in students to challenge these models, as it is typically an affront to their patriotism that the United States is nearer the middle or bottom of the developed world, respectively, in the two measures. Students reflect on the Human Development Index as a model of well-being within a country at the end of the activity.

The first explicit modeling activity involved evaluating the amount of water that a student used directly and indirectly during a week, based on a self-reported inventory and statistics about the amount of water used in the production of food and the per capita water used in producing electricity and other goods and services in the United States. This activity also compared the energy inputs of the food that students have eaten to a subsistence diet. Calculations were done

with an annotated spreadsheet provided by the instructor. Following the activity, students were asked a series of reflective questions about the models:

1. What variable in the model do you think should be removed and why?
2. What variable in the model do you think should be added and why?
3. Is there a variable in the model that you agree is important, but disagree about the equations used? Explain. Would it be possible to verify the accuracy of this number? How?

The next modeling assignment involves evaluating various carbon footprint models. Students were first asked to brainstorm a list of variables that they believe would contribute to their carbon dioxide production. They also identified which factors they thought would have a larger effect. After going home and getting personal energy use data, they then were asked to examine at least three of the interactive carbon footprint models available on the Internet. These were models from different companies such as the Environmental Protection Agency, British Petroleum, and the Nature Conservancy. The students were asked to list the variables that each site used in its calculations and compare them to the other sites and to the student's initial brainstorming activity. Students were to pay particular attention to variables that were unique to or conspicuously absent from a site and to reflect on how these differences could be accounted for in another way. If these variables were not accounted for, what effects would this variable have on the accuracy of the model? A whole-class discussion of the models followed, with the instructor/researcher pointing out specifically one model that functioned differently from the rest. The Nature Conservancy model appeared to start from a national CO_2 emission average (from all sources, personal and commercial) and calculated deviations from that average based on self-reported tendencies such as the use of fluorescent bulbs or recycling. All of the other models

calculated forward from kWh of electricity used, miles driven, etc., but only calculated personal use. The contrast of these models addressed the idea that model are intentionally created for different, specific purposes (such as showing direct carbon emissions only vs. showing direct and indirect carbon emissions) as well as trade-offs of complexity vs. accuracy and inclusion/omission of particular variables. The final requirement was for students to create their own carbon footprint model using the variables they believed were important. Through this learning experience, the students should identify that there were multiple models for the same concept, and that it was the creator of the model who ultimately decides which variables to include.

The next modeling assignment involved using the PhET greenhouse effect model and the Java Climate Model. Both were used to predict the relationship between greenhouse gas emissions, greenhouse gas concentrations, and global warming. Students were then asked to manipulate various variables of the model and to assess how these changes affected the output variables. They also were asked to reflect on the underlying assumptions, such as how gross domestic product, energy use, and population were assumed to change over time for the model. Students were also asked to look at and to comment on the various assumptions of this and other global warming predictions, and also to comment on the significance of the fact that the nine major global warming models each gave distinctly different predictions for the temperature increase by 2100 AD.

For their final modeling project, the students were required to create a model based on the research they conducted on a particular environmental issue. For example, was the energy used in making ethanol greater than the energy of the ethanol produced? Students selected variables that they felt were important to answer the question, and proposed how these variables could be

related. Students submitted three drafts of the model. Draft one was a qualitative model, identifying necessary variables and proposed linkages. Draft two was a spreadsheet model, submitted for feedback. Draft three was the final version. Students then wrote a short paper (approximately three pages) reflecting on the construction process and the strengths and limitations of their model, and their attempt to use the model to test a hypothesis.

Copies of all assignments are located in Appendix D.

Variables and Definitions

Variables.

The score on the Classroom Test of Scientific Reasoning is the independent variable. This pretest score is interval level data that was initially to be analyzed at the ordinal level of reasoning level (pre-formal, transitional, full formal, and post formal). Ultimately, the analysis chosen treated it as interval level data.

The dependent variables include a number of different sources of information. There are scores on two different tests as well as data from a number of student assignments as well as a final project, all of which was quantified. Additional information was collected from email, video recordings, audio recordings, and interviews, and some of this information was quantified.

One of the two primary dependent variables was the score on the Student Understanding of Science and Scientific Inquiry Questionnaire (SUSSI). A modified version of this test, which contained both Likert-scale and free-response questions, was used as both a pretest and a posttest. The free response portions were scored with a rubric tested on peers of the students of this study, and scored at least twice, including at least once by someone other than the author. The scores on this are interval level data and were analyzed as such. This score was the primary quantitative measure of nature of science knowledge.

The second major dependent variable was the score on Student Understanding of Models in Science (SUMS) instrument. A modified version of this Likert-scale test was the primary quantitative measure of modeling knowledge. The modifications included language changes to reduce ambiguity, reversal of some answers (every single question originally had as its best answer *strongly agree*, so the reversal of some questions so that strongly disagree is as likely to be the best answer as strongly disagree seemed appropriate), and the addition of free-response questions. This variable was interval level data.

While these variables were the variables among which the researcher wished to find relationships, other variables may have also been able to explain the observed data. As in most educational research, the list of potential extraneous variables was quite large. These extraneous variables were collapsed into two major groups.

Since the dependent variable was gain in a measure of academic knowledge, the first category of extraneous variables are those that would influence a student's success at learning new material. Study skills, time and opportunity to study, willingness to attend office hours and otherwise seek appropriate help when needed, and reading comprehension are all variables that might influence how successful a student was at learning new material. Based on this assumption, a student's current grade point average (GPA) should reflect of a combination of these variables. Thus, GPA was recorded and considered as a possible covariate/cofactor for data analysis, despite the fact that it was not significant in the pilot study.

A second category of variables that might influence students' performance consisted of demographic variables. These variables include age, gender, ethnicity and socio-economic status. Of these, age and gender would be of the most concern because Adey and Shayer (1990) found an interaction between age, gender, and developmental level. During the pilot study for

this study, a tentative relationship between scores on the CTSR and gender was uncovered. Therefore, these demographic variables also were recorded for possible use in data analysis.

Operational definitions.

The following definitions are used in this paper:

Developmental Levels. Two of Piaget's cognitive developmental levels, and one further level, are of interest in this study.

Concrete operations. The student is able to conserve and reason spatially, as well as do arithmetic with numbers that do not specifically represent concrete examples. However, the individual still has difficulty with abstractions. While this stage may appear in children as young as seven, it tends to appear in pre-adolescence (Piaget & Inhelder, 1955). This may be the terminal stage of development for some people. Operationally, CTSR scores < 14.5 are indicative of a student at the pre-formal or concrete operations level.

Formal operations. Individuals who reach this stage are able to reason formally, including performing such tasks as compensation isolation of variables, and systematically formulating and testing hypotheses. While Piaget himself observed this stage beginning as early as the onset of adolescence (age 12), evidence suggests not all students reach this stage, and certainly not by college (Lawson et al., 2007; Adey & Shayer, 1990). Operationally, CTSR scores above 14.5 (but less than 20.5 if post-formal reasoning is included) are indicative of a student at the formal operations level.

Post-formal reasoning. Lawson et al. (2007) has found a post-formal stage that involves the ability to reason hypothetically when the variables (agents) responsible for effects are themselves unseen. When and if it appears, this post-formal level is typically found in students over the age of 18, and this stage is operationally defined by a CTSR above 20.5.

Level one models/modelers. Level one modelers see the purpose of models as being an exact replica of the target (most typically an object), the purpose of which is to allow the users to accurately show or see what the object looks like. Students and answers conveying more of this conception of models (as opposed to the level two or level three definitions) will be labeled as level one.

Level two models/modelers. Level two modelers see the models as having deviations from the target, usually to make some aspect of the target clearer. These deviations make the model more useful for communicating about the target by, for example, leaving out an unnecessary aspect, emphasizing an important aspect or representing an imaginary concept more concretely (such as drawing an equator line on a map). Level two modelers are more likely to consider models of processes or systems in addition to models of objects. They see the purpose of models as being communication. Students and answers conveying more of this conception of models (as opposed to the level one) and which do not demonstrate the level three conception of models to form hypotheses, will be labeled as level two.

Levels three models/modelers. Level three modelers see models as representing the behavior of a target, which may be an object, process, system or other phenomenon. In addition to level two conceptions regarding the purpose of models and need for the model to differ from the target, a level three modeler is characterized by using the model's ability to accurately represent the behavior of the target to make predictions about the target, if some aspect of the model is changed. Students and answers conveying the idea that models are used to hypothesize will be labeled as level two.

Definitions

Inquiry. Inquiry is an approach to teaching and learning where students learn by first interacting with data in order to develop concepts. This approach is exemplified by the 5E model of Trowbridge, Bybee, and Powell (2000).

Model. As indicated previously, this is the central definition to the study. For this study a model was defined as a representation (physical, conceptual, or mathematical) of a target phenomenon, intended to communicate significant aspects of the phenomenon and to form hypotheses about the phenomenon.

Mathematical model. A mathematical model is a model as described above, where the phenomenon typically represents a system and its component parts that can be defined by variables and are quantifiable. As mathematical models were the primary models being analyzed and constructed, the mathematical models were further expected to accurately show the relationship between variables in the system. Within the constraints of a mathematical model, the hypotheses students formed took the form of effects shown when certain changes to the system were made.

Scientific model. A scientific model is a hypothetical-deductive model with unseen causative agents.

Data Collection

Quantitative instruments.

The Classroom Test of Scientific Reasoning, CTSR, (Lawson, Alkhoury, Benford, Clark & Falconer, 2000; Lawson, Banks & Logvin, 2007) was used to assess cognitive developmental level (the independent variable). This particular version of the test includes not only questions to discern what Piaget called pre-formal or concrete operational thinkers from formal operational

thinkers (Piaget, 1977) but also contains questions to assess students' ability to appreciate invisible causal agents. The first version of this test (Lawson, 1978) was one of the first reliable group administered test of developmental level, and its validity has been well established over the ensuing three decades. Subsequent versions of this test have evolved so that in its current form, it is longer (20-26 questions depending on the version), completely pencil and paper, and now measures Lawson's proposed fifth stage of development in addition to determining whether students are concrete, transitional, or formal. The Cronbach's alpha reliability of these exams varies from .79 (Lawson et al., 2007) to .81 (Lawson, Drake, Johnson et al., 2000). The version used in this study has 24 questions, four of which assess the post-formal stage. Students were awarded one point for each correct answer. Table 4 gives information regarding the scoring and interpretation.

Modeling ability was assessed by a modified version of Students' Understanding of Models in Science (SUMS), (Treagust et al., 2002). This 27 question, Likert-scale survey aims to assess several aspects of modeling knowledge with separate sub-scales. These sub-scales are: (a) multiple representations of scientific models; (b) whether or not scientific models are exact replicas; (c) the explanatory nature of models; (d) the use of scientific models; and (e) if, how, and why scientific models change (Treagust et al., 2002).

Cronbach's alphas for each of these sub-scales ranged from 0.71 to 0.84, indicating a high degree of self-consistency and thus reliability of the data. Item to total correlations were above .45 for 26 of the 27 items, and a bi-variate correlation of each sub-scale was significant at the .01 level. There was no validity measurement reported for this instrument, perhaps because no established, valid criteria or scale exists for measuring modeling ability exists. As the seminal work in modeling, Grosslight, Unger, Jay, and Smith (1991) arrived during a shift

towards increasingly qualitative research in education, this lack of scales and validity makes some historical sense.

The SUMS instrument was developed with students younger (age 13-16) than the students in this study (age 18+). After administering this test in a pilot study to peers of the students that were studied, several changes were made during and after the pilot study. Perhaps because of the difference in age or dialect (Australian English vs. American English) the wording of several questions was clarified (see Appendix A) when difficulties appeared in follow up interviews with student understanding of what the question was asking that obscured an accurate measure of whether or not the student understood the science concept involved. One further revision was made because of a large concern for a potential risk to the validity of the instrument. Almost all questions were phrased to the positive; such that the most correct response, the most likely response (in the case of misconceptions) or both was strongly agree. This uniform wording is perhaps why there is such a high correlation between sub-scores, as most respondents answered agree to most answers. In the pilot, as in Treagust et al. (2002), the overwhelmingly most frequent response was agree. With so many questions with the same answer in a row, during the pilot study concerns were also raised that the students were no longer carefully reading the question and instead were answering more or less reflexively. It appeared necessary to reverse the wording of approximately half of the questions to separate the conscious agrees from the reflexive agrees. If the same pattern of agrees remained after rewording, there is a validity issue with the instrument. If, as is anticipated, the pattern changed to fit the reverses in wording, then the test is more likely to be acceptable.

One aspect of modeling that is not addressed by this SUMS test is the construction of models themselves. While the if, how, and why scientific models change sub-scale measures

Table 4. *Interpretation of the Classroom Test of Scientific Reasoning*

Score	*Piagetian Stage*	*Lawson Stage*	*What students can do*
0-8	Concrete operational	Level 3	"Not able to test hypotheses involving visible causal agents"
9-14	Formal operational	Low Level 4	"Inconsistently able to test hypotheses involving visible causal agents"
15-20	Formal operational	High Level 4	"Consistently able to test hypotheses involving visible causal agents"
21-24	--	Level 5	"Able to test hypotheses involving unobservable entities"

From (Lawson, Alkoury, Benford et al. (2000), p. 1004).

ideas related to the modification of models, which ought to be similar to ideas concerning construction of models, this aspect of the study is significant enough that it deserves its own questions. Three out of seven of the characteristics of models involve model design are also discussed in Van Driel and Verloop (1999) including (a) some aspects of the phenomenon intentionally excluded from the model for simplicity's sake (b) conscious choice in the selection of factors to include and exclude, and (c) development through an iterative process including comparison to empirical data. Justi and Gilbert (2002a) also mention observations and data sources as the start of model building, and the importance of testing with empirical data. Perhaps the key aspect of a scientific model is its ability to generate a testable hypothesis model (Van Driel & Verloop, 1999). Thus, the following questions were added to the pretest.

12. A headline reads "Global warming model predicts that sea level WILL rise 2 meters by 2100 AD". a) What do they mean by *model* and b) how was this model created?
13. What is the most important characteristic of a model?
14. List as many science models as you can think of.
15. Multiple models exist of the same phenomenon, such as a map of the United States. Why?

The first question (#12) attempted to address model construction and to determine if students recognized a mathematical model as a model. While students may not have been able to name a mathematical model when asked (in question 14), could they recognize a mathematical model, when given one? This question also assessed the choice of factors (in this case, variables and relationships between them) in model construction. While this question was good as a pretest question, since this topic was studied in class, a conceptually similar yet not explicitly studied question "A headline reads 'Economic model predicts that China's economy will pass the U.S.'s economy by the year 2025'. What do they mean by *model* and how was this model created?" seemed more appropriate for the posttest, although its use created instrument validity concerns. Because the replacement question is similar in form, but unique in content from the original question and more importantly, from the specific content examples used in class, the posttest was given the new question on an economic model instead of pretest question on the global warming model.

The second question (#13) attempted to classify students along the three levels of Grosslight et al. (1991). A response that stated that the purpose of a model was *to show* indicated a level one modeler who was focused on surface similarities between the model and the phenomenon. A response that stated that the purpose of a model was *to explain* indicated a

level two modeler, who may see models as a way to aid communication about an issue, including processes as opposed to merely physical features. A response that stated that the purpose of a model was *to test* a hypothesis indicated understanding at the third level, consistent with scientific understanding of models.

The third question (#14) provided further data on students' conceptions of models beyond the physical, and students received one point for each distinct type of model (conceptual, physical, mathematical) that they named.

The fourth question (#15) provided an opportunity for students to reflect on the purposes of multiple models of the same target, in a context that should be familiar to more students (students should have been exposed to topographic maps, geopolitical maps, road atlases, etc. in their K-12 education and everyday life.). While a science example would have been desirable, interviews during the pilot could not reveal any science phenomenon where the context did not prevent some students from being able to answer correctly because they were totally unfamiliar with the concept, let alone the models.

One final modification was the transferring of the test into the Desire 2 Learn classroom management system for ease of data analysis and improved reliability of grading.

For a complete version of the final, modified SUMS instrument, see Appendix A: Instruments.

Student Understanding of Science and Scientific Inquiry questionnaire (SUSSI) was used to assess nature of science knowledge (Liang et al., 2006). This questionnaire consisted of six distinct sub-strands, each designed to measure a specific aspect of the nature of science knowledge. Each sub-strand consisted of four Likert-scale questions followed by a free-response question, and these questions were modified during and after the pilot (see Appendix

B: Achieving inter-rater reliability). These sub-strands were: (a) tentative nature of science knowledge; (b) observation and inference; (c) the scientific method; (d) creativity vs. rationality in science; (e) scientific theories and laws; and (f) cultural and social factors influencing science

The Cronbach alpha was 0.67 for the sample of students tested in the United States. This instrument provides no objective validity score. Again, it may be that there is not an agreed upon scale or criteria for rating nature of science knowledge, since both the multiple choice researchers (Aikenhead & Ryan, 1992) and the free-response researchers (Abd-El-Khalick, Lederman, Bell & Schwartz, 1998) consider the other group's methodology flawed. However, both the VNOS (Abd-El-Khalick et al., 1998) and the VOSTS (Aikenhead & Ryan, 1992) which the SUSSI is based on, have been extensively used, and have had face, content, and construct validity established through reviews by groups of nature of science experts. Validity, while not measured, was claimed to be improved by revisions after comparing the results of semi-structured student interviews, considered the most valid NOS assessment, to student answers. The SUSSI, which draws from both the VOSTS and the VNOS, was further checked for validity with a panel of nature of science experts and through multiple revision cycles.

However, this study only attempted to address two or at most four of these subscales; (a) student's understanding of theories (through analogy with models) (b) the tentative nature of these theories (revision and modification), with possible (c) understanding of the role of creativity in science (which variables to include), or (d) the scientific method (construction of models, although not a typical experiment, is an appropriate scientific method of investigation, particularly for phenomena that may not be observed directly). A pilot study revealed that students' views on observation and inference remained largely unchanged (as might be expected from a class that did not contain an observational/experimental component) as did their views on

the cultural and social factors. Therefore, these questions were dropped in the interest of time, reducing the overall length of the instrument and reducing the likelihood of test-fatigue which was reported in the pilot study interviews as a reason for less thoughtful answers.

These questions, too, were moved to the Desire 2 Learn classroom management system. See Appendix A: Instruments, for a complete list of questions.

Qualitative measures.

In addition to these quantitative sources, qualitative sources of information were used to attempt to explain the numerical results. Some of these sources were more structured and intentional *a priori*, while others looked at emerging trends in the data.

A subset of students was interviewed following the pretest and posttests in an attempt to determine if the written instruments (SUSSI and SUMS) accurately gauged student knowledge. This subset of students was selected from the available volunteers to represent the best range of student scores. Prior to the pretest interview, student responses on the SUSSI and SUMS were analyzed as a basis for the interview. The interviews themselves proceeded through several stages and were audio recorded. First, any ambiguity in the free response portion of the instruments was clarified. An example of ambiguity would be if a student said that a model is not an exact copy of a phenomenon, but did not explain why. If, upon further probing, the student explained that he made the statement because a model in no way resembles the phenomenon, then that student's score would be low. If, however, upon clarification the student indicated that a model is similar in some ways to the target phenomenon, but had some aspects emphasized and others deleted to make it easier to understand and use, then this student would receive a maximum score. Once these clarifications were addressed, other ambiguities of the pretest were probed in the following order. Second, inconsistencies between scores on the

Likert-Scale and free-response questions addressing the same sub-strand were questioned in an attempt to determine which better represents the student's true views. Third, Likert-Scale questions that were answered inconsistently from other questions in that strand, Likert-Scale questions answered two levels or more away from the mean ("agree" as opposed to "disagree," for instance), and Likert-Scale questions answered with a neutral response were probed when time allowed. During the pretest interview, no new questions were asked, but the students did have an opportunity to explain their answers on the pretest in more depth and ask for clarification or rewording. The interviewer also provided additional prompts to the interviewee in an attempt to elicit a better scoring answer. It was felt that these prompts were useful because students may not readily have examples of the concept in question, but were fully capable of analyzing a situation provided and thus able to demonstrate understanding of a concept. These prompts are listed in Appendix C: Interview Protocols.

A posttest interview was also conducted. These students were selected from the group of volunteers to give the most representative sample of gain on the SUMS and SUSSI instrument. As with the pretest interview, the first purpose was to clarify any ambiguity. However, a more important aspect of this interview was to ask students to compare their pretest and posttest answers from select questions. These questions were selected based on how much their answers changed from the pretest, in order of the greatest change, and explored as time permits. Students were asked to give insight into how and why their answer changed or did not change from the pretest to the posttest. The researcher looked for student statements pointing to specific parts of the treatment that caused a change in the student conceptions. Additionally, these interviews served as a check to the rigor of the study by looking for references to the activities vs. references to direct instruction. Direct instruction would damage the validity of the study on

two levels. The first danger would be if the student was only be parroting the answer the instructor said and did not understand or did not internalize the concept. The second, and more important, danger would be that this study purports to examine how students could gain understanding of the nature of science and of modeling through interacting with models, not by being told about them. Direct instruction in the nature of science, and to a lesser extent, the nature of models was held to a minimum. Both sets of interviews were recorded.

The individual student answers on modeling assignments constituted another important source of information. The follow-up questions to each assignment encouraged students to reflect on the variables selected, their purpose, how they were linked together, and what variables were omitted and why. The answers to these questions provided insight into the students' developing understanding of models. Two analyses were conducted on each reflection. First, a modeling score from one to three (Grosslight et al. (1991) to show, to explain, to test) was to have been assigned individually to each student. Second, as per Creswell (2003), data was read for emerging trends and then re-read to code such trends. A priori, some expected trends that were coded were keywords associated with each of the levels. Words such as *copy*, *show*, *exact*, and phrases indicating that the overall goal of a model is a completely accurate, static copy of reality were coded as level one, a naïve understanding. Phrases indicating multiple models, models to communicate ideas, models as approximations or simplifications of reality, or modifications of models received a score of two. Phrases indicating models being used to create and test hypotheses received a score of three. Other trends and codes were analyzed as they emerged.

The capstone task in this study involved constructing a mathematical model of a phenomenon that could be described, at least in part, through chemistry and required students to

think in variables, symbols, and equations about objects (atoms, molecules) that they have never directly observed. A student at the concrete operational stage of development was expected to have difficulty with this task, whereas a student with formal or post-formal operational development was expected to have more success. The final modeling project was scored like the reflections above, first on the level of modeling (one through three) and second on emerging trends. Third, the final model was assessed as to the two levels of causative agents as described in Lawson, Clark, Cramer-Meldrum et al. (2000) and Lawson, Drake, Johnson et al. (2000). A coding of level one was given to models involving concrete agents. A coding of level two was given to models involving unseen agents. A coding level of three (for a new, hypothetical construct) would have warranted a score of three, but was necessary.

In addition to the structured information in the tests, assignments and final project, some unstructured information was also recorded. Audio-recordings of office hours during which modeling or the nature of science was discussed and all emails regarding modeling and video-recording of the class modeling sessions comprised the unstructured information. The video-recording of the class sessions primarily was used to verify implementation validity of the instruction and assure that explicit teaching to the test did not occur. Other sources of information were examined and coded in the event that they may provide additional richness to the reflections. These other sources of information included student classwork, test answers not directly related to modeling and student comments in class or office hours. Excerpts from assignments were copied and retained. Relevant comments were recorded as soon and as accurately as possible. When possible, the quote was checked against the recollection of the speaker.

It was observed during the pilot that students sometimes gave answers indicating a deeper understanding of an aspect of modeling as part of one answer to a question focusing on a different aspect of modeling, but then did not repeat this answer in the question that was specifically trying to assess this concept. While credit for these misplaced answers was not given in the strictly quantitative analysis of the individual questions involved, these answers can be recorded and their frequencies tabulated for a test wide statistic (although in reality these did not differ from the individual question statistics).

All of these sources of information were used to support the quantitative data.

Sample

The sample consisted of four sections of Chemistry 304, the Environment and You, taught by the researcher during Summer and Fall Semesters 2008. These classes had the following characteristics. Each section had a maximum enrollment of 30 students per section. It was upper division – typically sophomores or above. It was an outer cluster class – a class designed to integrate and build upon aspects of the inner cluster (math, reading, writing, and critical thinking) and middle cluster (in this case, primarily natural science) classes that students have taken prior to this class. It carried a Learning Area 10 (People and the Environment) designation – according to the Dragon Core liberal studies program, this class does not count as a natural science course, but as a course in the area of People and the Environment. It carries liberal studies designation– these students were typically not chemistry or even science majors. It was writing intensive – students completed 16 pages of formal writing using multiple drafts and revisions. Finally, it carried a chemistry prefix rubric – chemistry is the subject matter about and through which the above writing and liberal studies goals were addressed.

The sample seemed to be representative of the students at the university in terms of ability, history, gender, and major distribution. The following describes the student body, with all numbers reflective of data from the four years previous to data collection. The university is primarily undergraduate (7,100 undergraduate students, 400 graduate students), female (60%), and white (only 5.2% of student identified themselves as a racial or ethnic minority, and 3.6% of students are international). Over 90% of students come from the two states closest to the university. Transfer students (40%) and students over 25 years-of-age (15%) make up a larger proportion of students than are found at some universities. Finally, the school is minimally selective, admitting 80% of applicants, who, on average scored a 21.7 on the ACT test and who are only slightly more likely than not (60%) to have finished in the top half of their high school graduating class Gill (2007).

Several relationships were explored during the pilot study. From this data, raw gains in pretest and posttest modeling and nature of science scores were examined for significant gain. Students showed a raw gain of 1.5 standard deviations, with a pretest to posttest correlation of 0.55. Sub-scores for *model construction*, *explanatory tools*, *use of models* and *changing nature of models* all had Cohen's *d* scores of 0.78 or greater, as did the most important aspect of the nature of science, student's conceptions of theories, laws and hypotheses. These gains appeared to be consistent with the small amount of quantitative literature that is available (Akerson, Abd-El-Khalick & Lederman, 2000; Sarri & Viiri, 2003). Although researchers like Lawson, Banks and Logvin (2007) showed correlations of 0.45 to 0.47 between scores on the Classroom Test of Scientific Reasoning and classroom performance (course grade, final exam scores), no correlations between the CTSR and gain were shown. Lawson, Alkoury, Benford et al. (2000)) also indicated that perfect correlation should not be expected, as, even though the study revealed

a correlation between increasing developmental level and the conceptual level of the task, neither all concrete students failed nor all post-formal students passed the balloon-transfer activity in their study, designed to measure abstract hypothesis testing. It was this use of correlation, rather than chi-squared or other categorical technique, that helped move the researchers to investigate other statistics.

Sample sizes were calculated based on a conservative 0.4 correlation between CTSR and gain on SUMS and SUSSI, which is in the middle of the "medium" classification for correlation (Cohen, 1988), and an effect size gain of 0.5 standard deviations between SUMS and SUSSI pretest and posttests, again "medium" by (Cohen, 1992). This gives a power of .804 with a sample size of 40.

Data Treatment

This study contains both quantitative and qualitative data, and the treatment for the quantitative data are described below (analysis of the qualitative data was described within the context of the description of each source). The following sources of data were used: (a) Scores on the CTSR pretest; (b) SUMS pretest and posttests; (c) SUSSI pretest and posttest; (d) classroom modeling assignments including the final modeling project; (e) interviews; (f) interactions caught on video and audio recordings. The handling of qualitative sources was discussed previously with each source, and emerged with the data. The handling of quantitative data is discussed below.

The CTSR, as a multiple choice test, was scored with one point awarded to each correct answer, summarized as follows: (a) concrete (0-8); (b) low formal (9-14); (c) high formal (15-20); and post-formal (21-24) (Lawson, Alkoury, Benford et al., 2000, p. 1004). Students also

were then classified according to these four levels, although the raw score was primarily used as the independent variable for analysis.

The Likert-Scale portions of the SUMS and SUSSI were scored according to the rubrics provided in Appendix B: Achieving inter-rater reliability. After the SUMS and SUSSI tests were scored individually and checked for inter-rater reliability, sub-totals for each strand and a grand total for modeling and nature of science were determined for each pretest and posttest. Normalized gain was calculated for each sub-score and the overall test. While a one-way Analysis of Variance (ANOVA) with four levels (concrete, early formal, formal, and post-formal) was to be used to determine if normalized gain for each sub-score and overall score was associated with developmental level at the alpha = 0.05 level, after consultation with Dr. Wendy Troop, an educational statistics consultant at North Dakota State University, a more appropriate statistic was chosen.

While the original analysis seemed good *a priori,* because a number of factors the analysis of the data using this approach was not successful. Poor design on some modeling activities prevented students from being able to demonstrate competency in the third level of modeling on some activities, resulting in ceiling effects. Many modeling activities had multiple subtasks, and in order to achieve a finer grain, the dependent variables associated with each subtask of a modeling activity was scored independently but simplified to a binary variable. While this could have led to the use of a chi-squared statistic vs. the four cognitive levels, with an *n* of only 60, a chi-squared with crosstabs resulted in too many empty cells and errors. Most importantly, while it appeared that many scores and gains were associated with cognitive development, they were better associated with the raw CTSR score than with the four levels. With the use of CTSR score (independent) as an interval variable instead of ordinal, and now

with only two options (correct or incorrect) for the dependent variables, the most appropriate statistic was the binary logistic regression. For all modeling activities, the binary logistic regression was thus used.

The analysis of the pretests and posttest was more straightforward. When measuring gains from the pretest to the posttest, normalized gain and Cohen's *d* were calculated for each of the sub-scores. Measuring the correlation between CTSR and gains on these sub-scores was slightly more complicated. First, the posttest scores were regressed onto the pretest scores. This provided an equation that gave a prediction for how much gain students should show from pretest to posttest. The student's predicted posttest score (or gain, it would amount to the same number) was then subtracted from the actual posttest score (or gain) to give a residual. A positive residual indicated that a student's gain was better than predicted, a negative, that the student's gain was less than predicted. A correlation was then performed between CTSR score and these residuals, to see if larger residuals (i.e. disproportionately large gains) were associated with larger CTSR scores, and vice versa.

CHAPTER FOUR

DATA ANALYSIS

The data on student conception of models and modeling consists of three main parts; scores on individual questions on the four small modeling assignments, scores on the final modeling project and associated assignments, and the scores on the modeling and nature of science pretest and posttest. A forth category of data includes variables about the students such as gender, CTSR score, and other potential covariates and factors. It is the results of this data about students that will be explored first, as this data will shape perception of the other data.

The sample

Sixty students participated in the study. Several pieces of data were collected from students, including gender, self-reported ACT score, self-reported grade point average (GPA), and semester the student participated in the study. In addition, each student's developmental level was measured using the Classroom Test of Scientific Reasoning (CTSR) (Lawson, 1978). Since developmental level is the primary independent variable in the study, these results will be examined first.

CTSR scores were normally distributed (Anderson-Darling Normality Test, $p = .431$, therefore not significantly different from the normal distribution) with all sixty students taking the test ($N = 60$, $M = 14.48$, $SD = 5.06$). When these CTSR scores were used to categorize students as high formal reasoning or above (CTSR > 14.5) vs. low formal reasoning or below (CTSR < 14.5),

31 students were classified at the high formal or above level and 29 students were classified at the low formal or below level.

The students came from four separate Chemistry 304 classes, with two classes each over two semesters (Summer, 2008 and Fall, 2008). Only 17 of the 60 students were from the summer classes, with 43 students coming from the fall classes. These two groups of students (summer vs. fall) did not have CTSR means that differed significantly $t(28) = -1.1$, $p = .28$.

The male students ($N = 20$, $M = 16.20$, $SD = 3.85$) who participated in this study were fewer in number but had significantly higher CTSR scores than the female students ($N = 40$, $M = 13.62$, $SD = 5.41$) who participated in the study $t(50) = 2.12$, $p = .039$. Although their CTSR score was higher on average, the self-reported GPA of the male students ($N = 20$, $M = 3.19$, $SD = 0.41$) was lower than the GPA of the female students ($N = 37$, $M = 3.31$, $SD = 0.46$) although not significantly so, $t(42) = -0.98$, $p = .34$, with 3 students not providing GPA data.

The relationship between self-reported GPA and developmental level was examined. No significant difference was found between the GPA of students at the high formal level or above ($N = 28$, $M = 3.33$, $SD = 0.47$) and students at the low formal level or below ($N = 29$, $M = 3.20$, $SD = 0.40$), $t(52) = -1.05$, $p = .30$. Furthermore, there was not a significant correlation between GPA and CTSR score $r = .17$, $p = .205$. CTSR score also failed to explain a significant proportion of variance in GPA, $R^2 = .029$, $F(1, 55) = 1.64$, $p = .205$. Thus, CTSR is not directly measuring the same ability as GPA, and the following analyses are not merely showing that students with a history of success (high incoming GPA) are merely continuing this success.

The remainder of this chapter will provide data relevant to the research question and sub-questions, repeated below.

Research Questions

Is attainment of the formal operational Piagetian level of understanding necessary for a model-based environmental science curriculum to increase students' understanding of models and the nature of science?

Sub-questions:

1. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations), and is that improvement related to Piagetian level?
2. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the utility of models (communication, simplification for study, prediction), and is that improvement related to Piagetian level?
3. Does a curriculum emphasizing student comparison, refinement, and creation of models improve student understanding of the relationship between models, theories, and the scientific method (models operationalize theories, allowing them to be tested with the scientific method), and is that improvement related to Piagetian level?

Pretest and Posttest Analysis

The differences between pretest and posttest scores were intended to be the primary quantitative measure of student gain in understanding of models and nature of science (NOS) across the course, and as such, represent the best overview of data relating to the research question and sub-questions listed previously. The only data not provided by the pretest and posttest was data on whether or not a student could construct a model, and that question will be examined in the next section using the data on the final modeling project.

The instrument used for the pretest and posttest was a modified version of the SUMS (Treagust, Chittleborough & Mamiala, 2002) combined with a modified version of the SUSSI (Liang, Chen, Chen, Kaya, Adams, Macklin, & Ebenezer 2006). This combined instrument contained Likert scale as well as free-response questions regarding models and the nature of science. As described in Chapter Three, Methods, each Likert-Scale question was valued at on one point and was scored on a five step scale, with one point awarded for the response most closely aligned with the accepted scientific view (strongly agree or strongly disagree in each case) and with 0.25 points deducted for each step away from the most scientific response to a score of zero for a response completely opposite of the scientific response. Each of the free response questions was scored on a three-point scale. This three-point scale served two purposes. First, where appropriate for the free response questions relating to modeling, these points roughly corresponded to the three levels of modelers in Grosslight, Unger, Jay and Smith (1991). In other words, a score of three represented a level three modeler, a score of two represented a level two modeler, and a score of one represented a level one modeler. Second, since the Likert-scale and free-response questions were combined into a single score, this heavier weighting of the free-response (three points) helped to balance the weight of the more numerous Likert-scale questions (at one point each).

The scores on the posttest (N = 60, M = 35.987, SD = 4.678) were significantly higher than the scores on the pretest (N = 60, M = 30.987, SD = 4.092), with a paired t-test showing $t(59) = -8.58$, $p < .001$. This difference was not only significant, but it was large, as it showed a 20.7% normalized gain and an effect size of d = 1.28, which Cohen (1998) classifies as large.

To determine individual student gain, posttest scores were regressed on pretest scores, yielding the following equation: Posttest score = 19.8 + 0.525 * Pretest Score, R = .459, R^2 = .211,

$F(1, 58) = 15.5$, $p < .001$. Hypothetical posttest scores were predicted using student pretest scores and the above regression equation. These hypothetical posttest scores were subtracted from the actual posttest score, giving a residual. These residuals were plotted against and regressed upon scores on the Classroom Test of Scientific Reasoning, yielding Figure 3, and the equation Residual = -5.473 + 0.375 CTSR total, $R = .456$, $R^2 = .209$, $F(1, 58) = 15.27$, $p < .001$.

A correlation of 0.456 is considered to be at the high end of the moderate range (.3-.5) by Cohen (1998). Overall, this data provides support of the research question that a moderate gain was both present and correlated to developmental level, but does not provide specific enough data to address the three research sub-questions. In order to examine these more specific relationships it was necessary to break the overall score pretest and posttest scores into sub-scores.

Questions were grouped into sub-scores, which will be examined in more detail in Appendix E. Sub-scores for various aspects of the nature of science and modeling were created by summing all questions pertaining to each particular aspect. For NOS questions, the original SUSSI instrument provided the structure regarding which questions to combine, and for modeling, the original SUMS instrument provided that structure. In a few cases with the SUMS, interaction with students in the follow-up interviews indicated that students perceived these questions in a way other than the author intended. Because this interpretation from the students appeared to fit a different, or even multiple, sub-scores, a few questions have been included in more than one sub-score. The sub-score categories are listed in Table 5.

Results and analyses performed.

The procedure of calculating residuals was again performed for each sub-score. Posttest sub-scores were regressed on pretest sub-scores. From this regression, hypothetical posttest sub-scores were calculated. Hypothetical posttest sub-scores were subtracted from actual posttest sub-

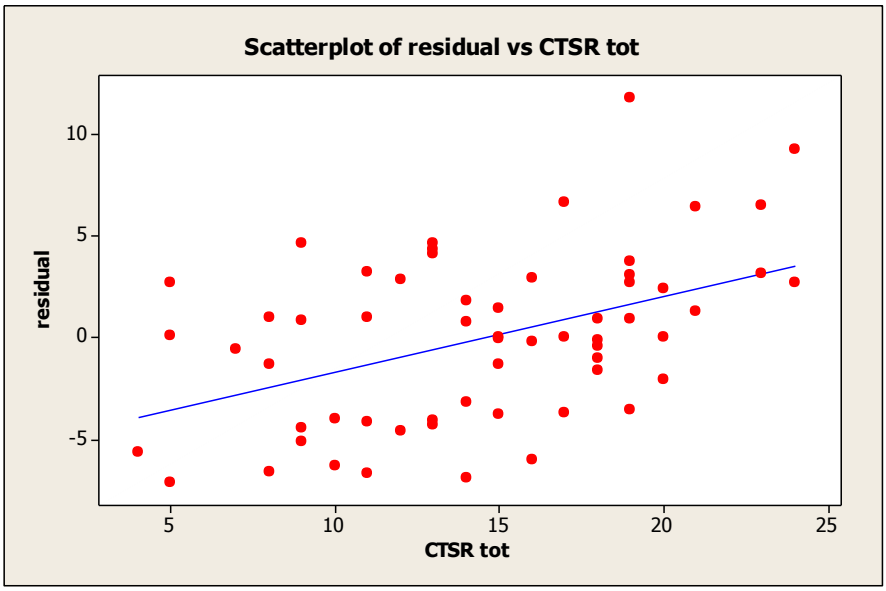


Figure 3. Scatter plot of residuals (posttest actual – posttest predicted) vs. CTSR score.

scores, to determine a residual. These residuals were regressed on and graphed versus CTSR. This analysis provided both an R-value and a p-value. The results for each of the sub-scores and the result for the test as a whole are presented in figure 4.

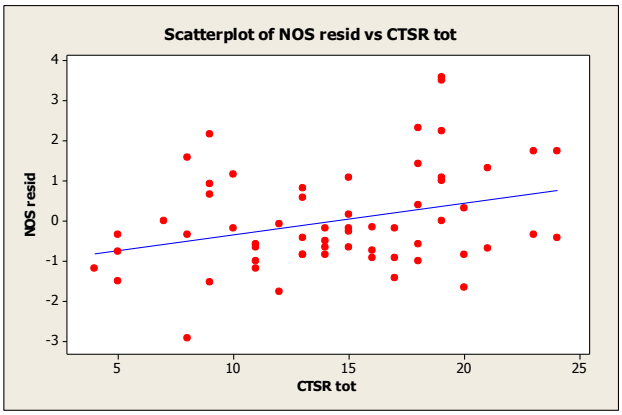
Question 39 does not appear in the above table as it represents a standalone misconception from the literature relating to multiple models and the educational construct of learning styles.

Overall, Table 6 shows that five of the 10 sub-scores (Nature of hypotheses, theories, and laws, Theory Change, Multiple Models, Exact Replicas and Use/purposes of scientific models) were all significantly (and positively) correlated with CTSR. All of these significant correlations except Exact Replicas could be classified as a moderate correlation according to Cohen (1998). The remaining sub-scores were not significantly correlated to CTSR.

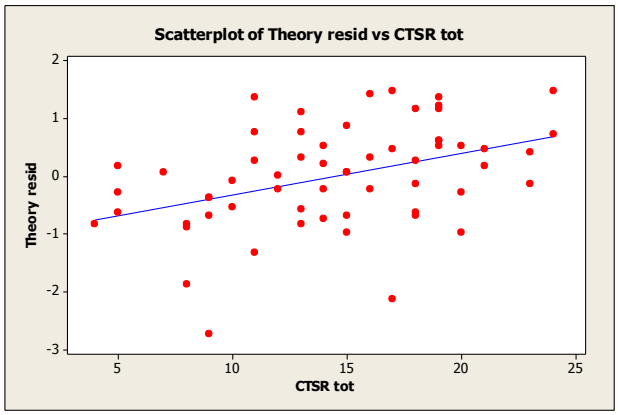
Table 5. Sub-score categories and component questions

Category	*Question(s) Likert*	*Question(s) Free-response*
Nature of science	1-5	6
Theory change	7-10	11
Multiple models	26, 27, 28, 29, 30, 31, 32	13
Explanatory tools	16, 17, 18, 21, 28	
Exact replicas	16, 19, 33, 34, 35, 36, 37, 38	
Uses/purposes of scientific models	13, 14, 20, 21, 22, 28, 29	13, 14
Changing nature of models	23, 24, 25	
Types of models		12
How are models created?	36	15
Scientific method(s)	40-43	44

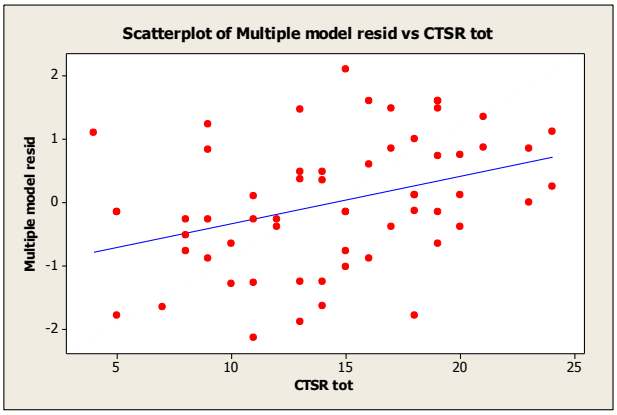
Nature of hypotheses, theories, laws vs. CTSR total



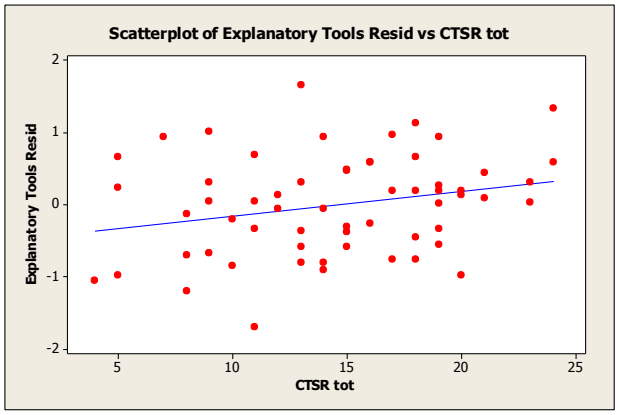
Theory change vs. CTSR total



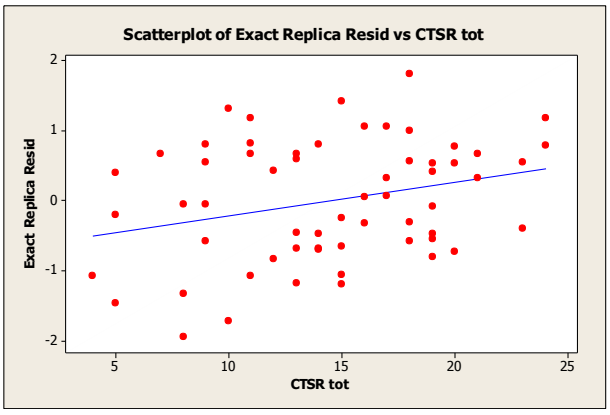
Multiple models vs. CTSR total



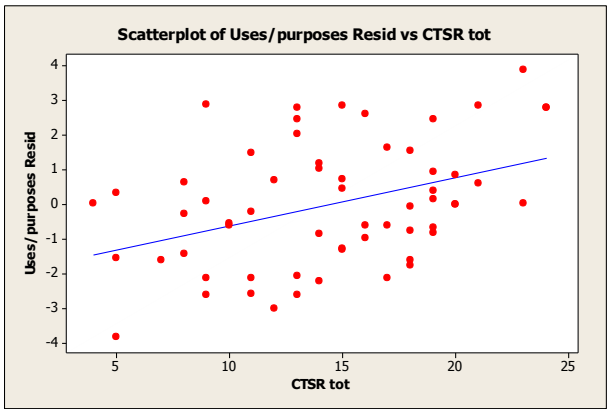
Models as explanatory tools vs. CTSR total



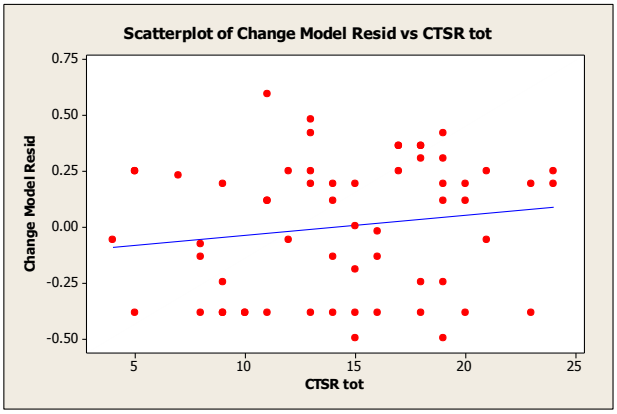
Models as exact replicas vs. CTSR total



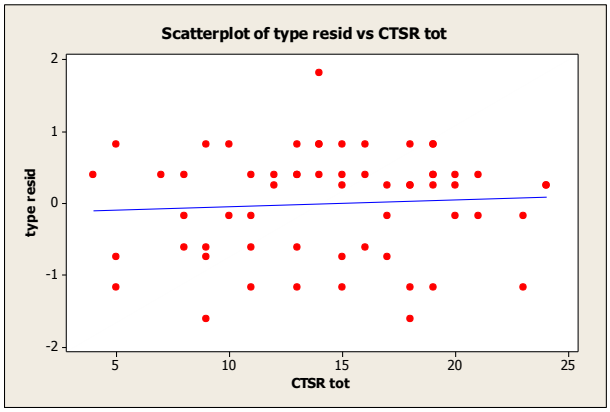
Uses of scientific models vs. CTSR total



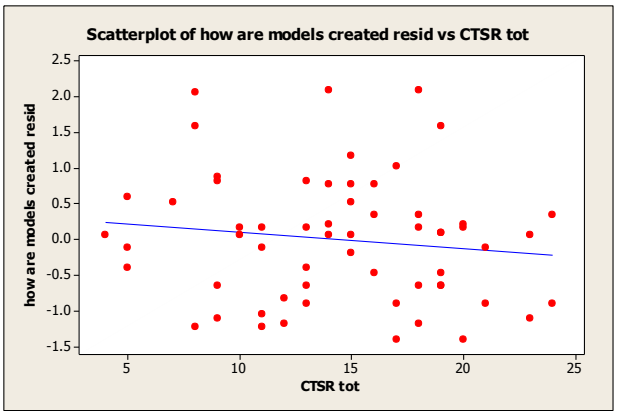
Changing nature of models vs. CTSR total



Types of models listed vs. CTSR total



How are models created vs. CTSR total



Scientific method vs. CTSR total

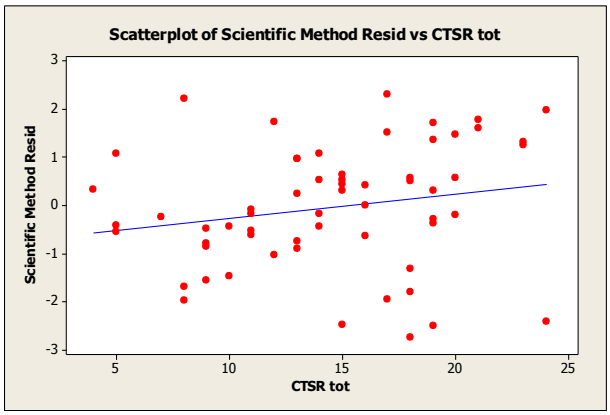


Figure 4. Scatterplots (with regression lines) of post-test residuals vs. CTSR total for each sub-score

Table 6. Correlation and p values for regressions of posttest residuals on CTSR for sub-scores and entire test.

Category	*r* for residuals regressed on CTSR score	*p* from a regression of residuals of sub-score on CTSR raw score
Nature of hypotheses, theories, and laws	0.3114 Moderate	0.015*
Theory change	0.4135 Moderate	0.001***
Multiple models	0.3674 Moderate	0.004**
Explanatory tools	0.2490 Small-moderate	0.055
Exact replicas	0.2863 Small-moderate	0.027*
Uses/purposes of scientific models	0.3838 Moderate	0.028*
Changing nature of models	0.1517 None	0.247
Types of models	0.1631 Small	0.213
How are models created?	-0.0574 None	0.663
Scientific method(s)	0.2 small	0.124
Total test	0.456 moderate	0.001***

* = significant at the $p = 0.05$ level

** = significant at the $p = 0.01$ level

*** = significant at the $p = 0.001$level

In addition to the above correlations between gain and CTSR, the size of gains for each sub-score was also calculated. Raw gain (posttest – pretest) was calculated for each sub-score and was converted into normalized change using (raw change)/(maximum possible change). Finally, raw gain was also converted to an effect size (raw gain/pooled standard deviation). Thus, both a measure of the size of the gain and the strength and of the relationship between the gains in each sub-score and CTSR score were achieved. These results are reported in Table 7.

All sub-scores showed a positive effect size, with several sub-scores achieving medium or larger effect sizes. Specifically, large effect size gains were seen in *multiple models* (1.40), *uses/purposes of scientific models* (1.33), *types of models* (1.15), and *how are models created?* (1.03). A large effect size (1.28) was also observed for the test as a whole. Medium to medium-large effect sizes were seen for *changing nature of models* (0.64) and models as *explanatory tools* (0.57). Only models as *exact replicas* (0.42); *scientific methods* (0.39); *theory change* (0.32); and *nature of hypotheses, theories, and laws* (0.13) showed less than medium effect size. *Scientific methods* showed the largest effect size gain of the NOS areas.

Although effect size is one way to measure the gain from pretest to posttest, it is limited by ceiling effects, since it is calculated from raw gain. Normalized gain provides a different measure of gain that does not suffer from the same ceiling effect problem as effect size. However, when negative normalized gains are present, a different metric (normalize change) can be used. In normalized change, instead of comparing both gain and loss to the maximum possible amount of gain, gains are still compared to the maximum possible gain, but losses are compared to the maximum possible loss. This caps the loss at -100%. The numbers for normalized change are listed in descending order here: *How are models created?* (43%), *types of models* (39%), *uses/purposes of scientific models* (39%), *changing nature of models* (37%), and *multiple models*

Table 7. Normalized change and effect size (Pretest vs. Posttest) by sub-score and total.

Category	Class average normalized change (pretest to posttest)	Class average effect size (pretest to posttest)
Nature of hypotheses, theories, and laws	-0.0132	0.1336 Negligible
Theory change	0.2437	0.3182 Small
Multiple models	0.3493	1.4027 Large
Explanatory tools	0.2230	0.5695 Medium
Exact replicas	0.2097	0.4224 Small-medium
Uses/purposes of scientific models	0.3884	1.3268 Large
Changing nature of models	0.3724	0.6448 Medium-large
Types of models	0.3917	1.1548 Large
How are models created?	0.4281	1.0348 Large
Scientific method(s)	0.1191	0.3913 Small
Total test	0.207	1.28 Large

(35%) all showed average normalized changes above 30%. *Theory change* (24%), models as *explanatory tools* (22%), and models as *exact replicas* (21%) showed normalized changes above 20%. In addition, the gains on the test as a whole revealed a 21% gain. Only *scientific methods* (12%) and *nature of hypotheses, theories, and laws* (-1%) showed less than a 10% normalize change. While this order is not exactly the same as for effect size, for both measures, when ranked from highest to lowest gain, the bottom five and the top five sub-score measures in both cases were

the same, thus providing a measure of triangulation in determining which sub-scores showed the greatest gain.

In conclusion, the data for the pretest and posttest indicated uneven gains across the various sub-scores, whether measured as normalized change or as effect size. These gains were sometimes significantly related to CTSR. As a whole, the pretest/posttest showed a large effect size and a significant relationship between normalized change on the test and CTSR score. Further analysis of this data is presented in Appendix E.

Small Modeling Assignments

Each of these four small modeling assignments had one to four embedded questions relating to the knowledge about models or modeling, including modifying a model, comparing models, using a model to make a hypothesis.

The human population lab.

In this activity, students were asked, "The HDI is a model used to predict how good life is in a particular country. … You are probably not used to this kind of a 'model' yet. But let us critique the model – what do you think about the inputs the creator used to arrive at this ranking? Do you think these are valid inputs/assumptions? Are there other assumptions you would include that they did not, if you were to rank the countries on their quality of life? Were there exceptions?"

The ideal answer should have stated that the student either agreed or disagreed with the HDI ranking, and thus the underlying model, and on what grounds. With regards to this analysis, whether or not they agreed or disagreed was irrelevant; it was the explanation which followed this answer which was relevant and thus determined their score. Furthermore, the student should have provided specific examples of either how a country had specific statistics that indicated a quality of

life at odds with its HDI ranking or examples of how this ranking was a good match to these statistics.

Almost every student (N = 55 of a possible 60) completed an analysis of the model and made comments regarding additional variables that they would like to see added into the model, and thus their answers were able to be scored. Five students did not complete this activity and their data was not included in the analysis. The majority (N = 45) of students agreed that this was an acceptable or mostly acceptable model, with four students explicitly disagreeing, and six students not explicitly committing to whether they agreed or disagreed with the model, but instead proceeded directly to analysis of the model.

Students answers were initially categorized as completely acceptable (N = 40, M_{CTSR} = 14.80, SD_{CTSR} = 4.54), completely unacceptable (N = 10, M_{CTSR} = 11.10, SD_{CTSR} = 5.13), or containing a mixture of both acceptable and unacceptable parts, i.e. partially unacceptable, (N = 5, M_{CTSR} = 13.60, SD_{CTSR} = 3.85). Note that the mean of students giving a partially acceptable answer falls in between the mean CTSR scores of the students who gave completely acceptable answer and those who gave completely unacceptable answers. The last two groups (completely unacceptable and partially unacceptable answers) were collapsed into a single group, (N = 15, M_{CTSR} = 11.93, SD_{CTSR} = 4.76) and binary logistic regression was performed (Minitab 16.1, 2010) z = 1.95, p = .051, with an odds ratio of 1.14 (i.e. for each point a student's CTSR score increases, the odds ratio that they will succeed on this task increases by 14%, exponentially. Another way to look at this is that the odds ratio for success nearly doubles with every 5 points of increase on CTSR). Thus, this data did not quite establish statistical significance at the $\alpha = .05$ level that CTSR score was linked to success critiquing this existing model. Examples of each of the answer

categories (completely acceptable, completely unacceptable and partially unacceptable) are contained in Appendix E.

Three possible further trends emerged, will be mentioned briefly here, and discussed further in Appendix E. Each of these trends relates to a major understanding of how scientific models are made and used. In every case the mean CTSR score of the group who did make these additional observations or comments was higher than for the students who did not make these observations or comments. No statistical significance could be determined for any of these trends due to the small number of students making these comments; however, these data indicate further exploration may be warranted. Twelve students checked the HDI model against data (M_{CTSR} = 15.25 for those who did vs. M_{CTSR} = 13.67 for those who did not), six students discussed issues of the accuracy vs. complexity tradeoff in the model (M_{CTSR} = 14.33 for those who did vs. M_{CTSR} = 13.98 for those who did not), and two students explicitly discussed the purpose of the model (M_{CTSR} = 19).

The resource lab.

The Resource Lab was more focused on the underlying model than the Human Population Lab, which focused more on class content. Prior to the Resource Lab, students recorded their personal food and water use for one week. At the end of the week, students entered their data into the spreadsheet model provided. Direct and indirect use of grain and water was calculated. In the case of food, the indirect use consisted of converting all food (and even some non-food items such as corn ethanol) into an equivalent quantity of grain. At the end of the Resource Lab, students were asked a number of specific questions relating to the underlying model:

1. List at least one factor in the model (from either the total water usage calculation or the total grain usage calculation) that you think you would delete. Why is this factor

unnecessary/wrong? How does deleting it make the model better? Does it make the model simpler? Do you think it makes the model more valid/accurate?

2. List at least one factor (in either the water or grain parts of the model) that the designers of the model did not take into consideration. How would adding this factor make the model more accurate? Would the increased accuracy be worth the additional effort? Can you speculate on why the creators might have left this factor out (bias, agenda, simplicity, accuracy, or inclusion elsewhere)?
3. List at least one part (probably a way that something is calculated) of either the water or grain aspects of the model that you think is wrong. Why do you think it is wrong? Why did the model creator put the "wrong" factor in there (bias, agenda, using an average instead of a personal number, using a number [like pop cans] to take into account other factors [like garbage in general])? Where might you go to find the "right answer"?
4. How could a model like this be used to test or create a hypothesis regarding lifestyle/diet choices and food or water use? Give a specific example.

Fifty three students submitted an assignment. The seven students not submitting an assignment were not considered in this analysis. Table 8 provides a summary of the results.

Analysis of student answers to question one revealed that these answers belonged to 12 emergent categories (in addition to seven students who did not submit the assignment, and were thus excluded from the analysis). Table 9 shows the categories of responses, the individual CTSR scores of the students giving that response, and the mean of these CTSR scores.

Analysis required collapsing these categories into fewer categories. The small group that gave an acceptable answer to this question was collapsed with those who suggested an acceptable

Table 8. *Resource lab, results of statistical test by question vs. CTSR score*

Task	*Number of students*		*Mean CTSR*		*Test result*
	Successful	*Unsuccessful*	*Successful*	*Unsuccessful*	
Q1. Variable deletion	14	39	18.64	12.68	$z = 2.26$, $p = .024$*
Q2. Variable addition	42	11	15.12	10.64	$z = 2.45$, $p = .014$*
Q3. Variable change	43	10	14.95	10.90	$z = 2.14$, *$p = .032$
Q4. Hypothesis formation	28	25	15.58	12.52	$z = 2.20$, *$p = .028$

modification of a variable rather than deletion, since both represented a correct meaningful *change*, if not *deletion* ($N = 14$, $M_{CTSR} = 18.64$, $SD_{CTSR} = 3.63$). Furthermore, most students suggesting a meaningful change started by suggesting a deletion of a particular aspect, but correctly reasoned that the suggested change in this case would likely yield higher accuracy than a complete deletion, thus implying both an understanding of the question and consideration of the impact of variable deletion/change on accuracy of the model. The rest of the groups submitting an answer were collapsed together since these all represented answers that were not a correct change or deletion (N = 39, MCTSR = 12.68, SDCTSR = 4.57) . A significant relationships between CTSR and score on question one was found, with z = 2.26, p = .024, and an odds ratio of 1.17.

Table 9. *Resource lab, question one, and associated CTSR means*

Answer	*N*	*CTSR scores*	*CTSR Mean*
*Acceptable answer	6	15, 18, 18, 19, 23, 24	19.5
*Suggested changing variable rather than deleting	8	11, 15, 17, 18, 19, 19, 21, 24	18.0
**Had both acceptable and unacceptable parts	5	11, 11, 15, 18, 20	15.0
**Deleted essential part of the model with specific (but wrong) reasoning	6	5, 13, 16, 16, 19, 20	14.8
**Deleted essential part of the model, no reason given	3	14, 14, 16	14.7
**Deleted an essential part of the model because that aspect did not apply to them	2	14, 14	14.0
**Answer did not address the question asked	7	9, 11, 13, 13, 13, 15, 20	13.4
**Deleted an essential part of the model because of uncertainty	7	5, 9, 15, 8, 13, 13, 24	12.43
**Deleted essential use of grain/water because essential use should not be counted against user	4	8, 10, 12, 15	11.25
**Deleted “nothing” because they felt deleting anything would make the model wrong	1	10	10
**Said they would delete “nothing” with no reasoning	6	4, 5, 10, 14, 14, 18	10.83

Table 9. *Continued.*

Answer	*N*	*CTSR scores*	*CTSR Mean*
**Deleted an essential part of the model because it made their impact score large and they felt bad	4	7, 8, 11, 17	10.75

* = collapsed into a single category (acceptable *change*) for statistical analysis

** = collapsed into a single category (unacceptable *change*) for statistical analysis

Students found more success on question two with what variables to add and scoring was cleaner, with answers falling cleanly into correct and incorrect categories, examples of each can be found in Appendix E. Most students (N = 42, M_{CTSR} = 15.12, SD_{CTSR} = 4.68) correctly identified and supported the addition of a particular variable. Nevertheless, a smaller group of students (N = 7, M_{CTSR} = 11.43, SD_{CTSR} = 5.62) suggested variables that were not applicable. A third group of students (N = 4, M_{CTSR} = 9.25, SD_{CTSR} = 4.50) gave answers with both acceptable and unacceptable suggestions. The mean CTSR for students suggesting an acceptable variable is larger than for either of the other two groups containing at least partially flawed responses. If these two groups with at least some response errors are collapsed into a single group (N = 11, M_{CTSR} = 10.64, SD_{CTSR} = 5.10) and compared to the students answering correctly, the relationship between CTSR and the answer on question two is significant using binary logistic regression with z = 2.45, p = .014, and with an odds ratio of 1.22 (the odds of success on this question double for every ~3.5 points increase in CTSR score).

Likewise, most students (N = 43, M_{CTSR} = 14.95, SD_{CTSR} = 4.85) were able to suggest *acceptable* ways that individual calculations in the model could be changed in question three. The students whose answers were completely *unacceptable* (N = 4, M_{CTSR} = 9.50, SD_{CTSR} = 4.04) were

eventually collapsed with the six students whose *answer contained both acceptable and unacceptable components* ($N = 6$, $M_{CTSR} = 11.83$, $SD_{CTSR} = 5.53$) to give the group that was used in statistical test, ($N = 10$, $M_{CTSR} = 10.90$, $SD_{CTSR} = 4.89$). The results of the binary logistic regression were significant, with $z = 2.14$, $p = .032$, and an odds ratio of 1.19.

The answers of questions four were more balanced than the previous questions, with acceptable ($N = 28$, $M_{CTSR} = 15.58$, $SD_{CTSR} = 4.98$) and unacceptable ($N = 25$, $M_{CTSR} = 12.52$, $SD_{CTSR} = 4.72$) answers in roughly equal numbers. The difference in mean CTSR scores between those who formed an acceptable hypothesis and those who did not was significant when analyzed with a binary logistic regression, with $z = 2.20$, $p = .028$, and with an odds ratio of 1.15.

Throughout the Resource Lab, students had the opportunity to make other comments regarding other aspects of modeling, such as the complexity/accuracy tradeoff and the creator's purpose or bias in a model. The three students ($N = 3$, $M_{CTSR} = 14.67$, $SD_{CTSR} = 9.50$) mentioning the complexity/accuracy tradeoff with slightly higher mean CTSR scores than the 50 students who did not ($N = 50$, $M_{CTSR} = 14.16$, $SD_{CTSR} = 4.85$), and the eight student mentioning creator's purpose or bias ($N = 8$, $M_{CTSR} = 10.62$, $SD_{CTSR} = 5.83$) had lower mean CTSR scores than those who did not ($N = 45$, $M_{CTSR} = 14.82$, $SD_{CTSR} = 4.72$), although neither difference was significant $z = -1.93$, $p = .090$ and $z = 0.11$, $p = .913$ respectively. Further information is available in Appendix E.

The carbon footprint activity

The Carbon Footprint Activity involved the students working with models more deeply than the previous two activities. For this activity, the goal of understanding multiple models was nearly as important a goal for students as gaining content knowledge. This activity required students first to brainstorm about factors they thought would contribute to their carbon footprint (task one), then to collect data (electric bills, car make, model and miles per gallon, miles traveled,

etc.) related to carbon emissions (task two), then to input this data into at least three different carbon footprint models from various websites and compare and contrast these models (task three), and finally after an in-class tutorial on how to create formulas in Excel, to practice making a small carbon footprint spreadsheet model of their own (task four). One difference between summer and fall sections occurred here, as the summer section was able to spend a little longer (approximately one half to one full hour, depending on how quickly the individual student completed the other parts of the activity) on developing their model in task four, mostly due to the smaller class sizes completing the tutorial portion and task three much more quickly.

Most of the follow-up questions emphasized the modeling aspects of this activity. There were several questions that asked students to examine the multiple models of the same phenomenon critically. The questions were designed for students to demonstrate (a) That they understood that there could be multiple valid models of the same phenomenon, especially if each model had a slightly different goal, (b) That one of these goals could be accuracy vs. complexity, and (c) That no model is likely to include all aspects of a phenomenon. The questions specifically asked are included below:

1. What was the range of your results (low to high)? Why do you think there was such a range? What does that mean about these models? What does that mean about your carbon footprint? With such a large range, how can we use these models appropriately? Were your results in line with others who used models from similar sites (site #2 seems to always be low, or site #8 always seems to be high for instance?)

2. Accuracy/completeness versus complexity. One reason for multiple models of the same phenomenon is that certain models are more appropriate for a deeper understanding, where more accuracy is needed, and thus more complexity is required. Other times, a quick "ballpark"

estimate might be appropriate. For each of the parts below, do you think the listed aspect made the model more accurate? Was the change in accuracy appropriate given the change in complexity from adding/removing that variable?

2.a. What unique questions did each site ask you (or include in their model) that you did not have in your brainstorming list (Task 1)?

2.b. Which of the factors that you felt were very important (from Task 1) did this model not seem to incorporate?

2.c. Were there any factors that this website "lumped together" or used an average for? Why would they use an average?

3. After analyzing your sites and their models, compare and contrast. Could you say which is "better?" Which site would you use? I would say "it depends." Take AT LEAST 2 of the sites and say how and why you would use one in a particular setting, but another in a different setting.

The scoring for question one was as follows. A student received a score of two if the student correctly answered both *why* the results from the various models did not agree and *how* models could effectively be used if the models disagreed with each other so much. The correct answer to the first part (*why*) is because each model used asked for unique data and may have calculated the result differently or made different assumptions, and most students answered this correctly. There were several acceptable answers to the second part (*how*), such as (a) use each model for its apparent purpose (some asked questions that required utility bills, some only asked about energy reduction techniques in the house. If a student lacked the utility bills, using the second model would be more appropriate.), (b) use models to compare to the stated national average (since most models reported a national average, which itself varied from site to site, even if these sites gave different carbon footprints, they tended to be fairly consistent about whether the

student's carbon footprint was above or below average), (c) use models to see how various changes to lifestyle affect the carbon footprint (For example: any model which used the variable *diet* could be used to assess the change in carbon footprint in switching from an omnivore to a vegetarian or vegan diet. Although all models may give a different carbon footprint based on the particular variables that are included in that model, any model including the diet in a reasonable way should show a roughly similar decrease in carbon footprint for a similar diet change to any other model including diet.), or finally (d) to use whichever model seems to have the most reasonable approach (most models had a section that explained the calculations in detail). A student also may have spoken of a particular purpose of the creator in making these choices.

A student received a score of one if they correctly addressed one or the other parts of the question (as above) and omitted or gave an incorrect answer (see below) to the other part. All but one student in this group correctly answered *why* the models gave different answers, but not *how* to use them.

A student received a score of zero if they made a statement that contained neither explicitly correct nor incorrect parts.

A student received a score of negative one if they answered one half of the question incorrectly, but did not specifically give a correct or incorrect answer to the other part. For example, by far the most common misconception regarding the first half of the question was that more questions equaled a bigger carbon footprint, which was not correct in practice or theory. A common incorrect answer for the second half of the question was to average the output for the various models, but certainly if a student identifies that a website has a bias, or omits an important variable (such as the carbon dioxide emitted from burning wood for heat), it would be unwise to include the results from this model in an average of other, more valid, sites.

Theoretically, a student could have received a score of negative two, by answering both the *how* and *why* parts explicitly incorrectly, but this did not occur.

Four students failed to submit this activity and were not included in the analysis. There was some confusion during the summer semester regarding this assignment, and four students during the summer semester turned in incomplete work (some questions answered, not others. For these students (49, 50, 54, 58) questions where they gave answers were recorded and analyzed, questions where they did not were not used in the analysis. Thus, the total number of responses varies for the questions on the Carbon Footprint Activity. The results for question one of the Carbon Footprint Activity can be seen in Table 10.

As can be seen from Table 10, only nine of the students answering this question failed to find some success. The mean CTSR score for the students scoring a -1 or a zero were both approximately a full standard deviation below the mean CTSR score for students scoring a one or a two. When these four groups were collapsed into the two; the lower sets of responses (scores of -1 and zero) (N = 9, M_{CTSR} = 10.11, SD_{CTSR} = 3.02) together and the two higher sets of responses (scores of one and two) (N =43, MCTSR = 14.63, SDCTSR = 4.81), a significant relationship between the answer to question one and CTSR score was found: z = 2.37, p = .018 and an odds ratio of 1.24 when analyzed with a binary logistic regression.

The scoring for question two was as follows. A student received a score of two if (a) they correctly identified one variable that the models left out that the student had identified (in task one), (b) one variable that the models included that the student did not think of previously (in task one), and (c) one part of the model where an average was used. In addition, they needed to explain how the average was used. Finally, this question seemed provide an opportunity for students to explicitly discuss the accuracy/complexity tradeoff.

Table 10. *Carbon Footprint Activity, question one results and associated CTSR means.*

Score on question one	*N*	*CTSR Mean*	*CTSR Standard Deviation*
-1*	5	10.40	1.34
0*	4	9.75	4.65
1**	27	14.33	4.51
2**	16	15.13	5.39

* Eventually collapsed into a single group, those answering no part of question one correctly.

** Eventually collapsed into a single group, those answering at least part of question one correctly, and no parts explicitly incorrectly.

A student received a score of one if they correctly identified two of the three variable issues regarding models (listed above). In practice, it was always that students did not successfully deal with the issue of averages. This part of the question turned out to be difficult for nine students as will be seen by the scores below.

A student received a score of zero if they gave an answer that did not address the question, in other words, if the student did not correctly discuss the use of averages, nor were mismatches between the student's variable list and the model's variables identified.

The results of the scoring of question two are given in Table 11. As can be seen from Table 11, 75% of the students who turned in this assignment were able to answer this question correctly. Furthermore, there is almost no difference in the means between the 39 students who answered the whole question correctly, and those nine students who did not address why the variables used an average. These two groups of students were collapsed into a single group. Statistical analysis did not produce significant results between students answering the question

Table 11. *Carbon Footprint Activity, question two, results and associated CTSR means.*

Score on question two	*N*	*CTSR Mean*	*CTSR Standard Deviation*
0	4	11.50	3.87
1*	9	14.33	5.05
2*	39	14.23	4.64

* These two groups were collapsed in the final analysis.

at least partly correctly (N = 48, M_{CTSR} = 14.25, SD_{CTSR} = 4.67) and those who answered incorrectly when compared by their CTSR scores, with a binary logistic regression producing z = 1.12, p = .264, and an odds ratio of 1.14. Given the low failure rate on this question (only 4 students); even though the difference in means and odds ratio are relatively large, this difference is not statistically significant.

Question three was scored as follows. To receive a score of a one, the student needed to explicitly state how these different, multiple models could be used appropriately. The two expected correct answers would involve using models appropriate to the task (using a model that contained many modes of transport if one wanted to investigate the impact to the user's carbon footprint if the mode of transportation were changed) or the audience (younger children may benefit from a simpler model that focuses on simple questions such as "does your house have fluorescent bulbs?" rather than "how many therms of natural gas did you burn last year?")

To receive a score of a zero, the student must not have successfully discussed how multiple models could be used appropriately.

Results. The mean CTSR score of the students who answered the question correctly (N = 47, M_{CTSR} = 14.47, SD_{CTSR} = 4.89) is almost a standard deviation larger than the CTSR score of the

students who did not answer the question correctly (N = 7, M_{CTSR} = 12.00, SD_{CTSR} = 2.31), however, this result was not statistically significant when a logistic regression was performed, with z = 1.28, p = .201, and with an odds ratio of 1.12.

In addition to the answers specifically aimed at each of the three questions, many student responses touched on other aspects of modeling that were not being specifically addressed in each question. These areas include the students discussing the bias/purpose of the creator or of the model (N = 25, M_{CTSR} = 14.36, SD_{CTSR} = 5.71) vs. those who did not (N = 31, M_{CTSR} = 13.81, SD_{CTSR} = 4.12), those students who did discuss the complexity/accuracy tradeoff (N = 43, M_{CTSR} = 14.77, SD_{CTSR} = 4.82) vs. those who did not (N = 13, M_{CTSR} = 11.69, SD_{CTSR} = 4.33), and the students who spoke of the ability to use a model to make a hypothesis or otherwise reason about a topic (N = 15, M_{CTSR} = 14.87, SD_{CTSR} = 4.44) vs. those who did not (N = 41, M_{CTSR} = 13.76, SD_{CTSR} = 5.02). Although the differences in the means point in the direction of students with higher CTSR scores performing better on these tasks, only the relationship between CTSR scores and discussing the complexity/accuracy tradeoff was close to significant when a binary logistic regression was performed, z = 1.95, p = .052, odds ratio 1.15.

The global warming activity

The Global Warming Activity was more heavily focused on content, secondarily focused on showing models, and somewhat less focused on assessing students' understanding of models. Although students were asked to work with various models and look at a particularly nice schematic of the underlying relationships between the variables in one model, there were only two questions that specifically asked students to think about the models themselves. These were questions seven (part d) and eight repeated below.

7.d. Global warming skeptics will often cite this disagreement about the exact number [amount of temperature increase predicted] as proof that [the existence of] global warming [itself] is uncertain. The other way to look at this is that no matter how you calculate it, at least some global warming is predicted. **Comment!**

8. An explanation of the various scenarios can be found at: http://en.wikipedia.org/wiki/Special_Report_on_Emissions_Scenarios . Again, feel free to go to the original IPCC report…

Which of the scenarios do you feel is the most likely, based on the Human Population Lab earlier this semester, etc.? Support your answer. If not scenario A2, how would using those assumptions affect the global warming predictions from #7, (above), which are mostly based on a scenario A2 Earth…

The correct answer for question seven would be that multiple models may have small variations in the way that they calculate a particular output, depending on the variables included and the weight given to each variable. However, all of these models, no matter how they were calculated, predict an increase in temperature. The fact that they do not agree on an exact number does not invalidate them, and in fact, that all of them reach the same qualitative conclusion (global temperature will rise) using different methodologies lends more credence, not less, to the idea of global warming. The expected misconception would be the idea that there is only one right model for a given phenomenon, and therefore, if these eight models do not give identical answers, at least seven of the eight (if not all eight) are wrong.

The correct answer for question eight would be that according to the information presented earlier in the semester, scenario A2's assumptions about population growth are probably too pessimistic and scenario A1 is more likely. Because models based on A2 assume a faster

population growth, and because more people are likely to mean more pollution and therefore more warming, models based on the A1 scenario should provide a lower prediction for warming. Other answers that disagreed with A2's assumptions about technology, globalism, and international cooperation that likewise assessed the impact on global warming that changes to these underlying assumptions would make would be acceptable. Answers providing both an answer (agree/disagree) and a logical reason scored two points. Unfortunately, students who incorrectly agreed that A2's assumptions were reasonable could not effectively answer the question about what impact these changes on the assumptions in the model would have on the model, and their scores tended to hit a ceiling.

The results for question seven are detailed below. The student responses were rated according to the process described in the introduction to this section. Most (N = 43, M_{CTSR} = 15.40, SD_{CTSR} = 4.71) students agreed that the fact that there are minor differences in the outputs of the various models is not a reason to declare all models void, whereas four students (N = 4, M_{CTSR} = 9.25, SD_{CTSR} = 5.56) explicitly disagreed. Eight students (M_{CTSR} = 12.25, SD_{CTSR} = 5.65) gave answers that were so vague that they contained neither an explicit agree or disagree and five students did not submit the assignment.

The group (incorrectly) disagreeing and the group with answers which could not be scored clearly were collapsed into a single group, i.e. those not giving a correct answer (N = 12, M_{CTSR} = 11.25, SD_{CTSR} = 5.56). A binary logistic regression showed this group had significantly lower CTSR scores from the students who agreed that different answers did not prove some of the models were wrong, z = 2.34, p = 0.019, with an odds ratio of 1.19.

The results for question eight are detailed below. The student responses to question eight were evaluated as described in the introduction to this section, and the results were tabulated in Table 12.

The results for question eight showed roughly equal splits between the six groups of students summarized in Table 12. This split means that no subgroups were large enough for adequate statistical testing. Thus, groups were collapsed based upon correct analysis, with students giving correct analysis (regardless of whether they gave the correct answer) in one group and those students who did not give correct analysis in the second.

When these categories were collapsed, the results were significantly associated with CTSR score, with those students whose answer explicitly and correctly analyzed inputs (N = 23, MCTSR = 16.39, SDCTSR = 4.62) scoring better on the CTSR than those who did not (N = 32, MCTSR = 13.13, SDCTSR = 5.03) when a binary logistic regression was performed, with z = 2.25, p = .025, odds ratio = 1.15.

Final Project

Initial variables submitted.

One of the first steps students took in creating their own models was to submit to the instructor a list of variables that the student felt were appropriate to creating their models, after the students had researched the topic on their own. Scoring this artifact proved difficult, with a full discussion of this procedure in Appendix E. The initial scores were broken down as follows, and summarized in Table 13.

Students were scored as having *almost no relevant variables* if they had less than 10% of listed variables as relevant.

Table 12. *Global Warming Activity, question eight, results and associated CTSR means.*

Answer to question eight	*n*	*Mean CTSR*	*CTSR Standard Deviation*	*Score*
Disagreed, but did not state how prediction would be affected	13	15.31	4.35	-2
Agreed, but said nothing further or answer had logic flaw	7	10.86	3.85	-1
Answer was not able to be clearly scored	12	12.08	5.71	0
Did not turn in the assignment.	5	14.40	4.39	
Agreed and gave analysis	10	16.80	4.98	1
Disagreed, and correctly stated how prediction would be Affected	13	16.08	4.80	2

Students were scored as having *more irrelevant that relevant* if they had between 10% and 40% relevant variables, with more irrelevant variables than relevant. Students were scored as having *about equal relevant and irrelevant* if they had between 40% and 60% relevant variables, with irrelevant and relevant variables making up an approximately equal percent. Students were scored as having *more relevant than irrelevant* variables if they had between 60% and 90% relevant variables, with more irrelevant variables than relevant. Finally, students were scored as having *almost no irrelevant variables* if they had less than 10% of listed variables as irrelevant.

In conclusion, the above data hints that there are some links between cognitive development and the quality of the preliminary list of variables submitted by the students, as nine of the 10 students with more irrelevant variables than relevant variables had CTSR scores at the

low formal level or below (<14.5) whereas 20 of the 26 students with CTSR scores at the high formal or above (>14.5) had more relevant variables than irrelevant. However, when the initial ratios of relevant to irrelevant variables were collapsed into two categories (more relevant that irrelevant variables vs. equal or more irrelevant variables than relevant variables) and a binary logistic regression was performed, no significant difference was revealed $z = 0.86$, $p = .388$, odds ratio = 1.05. Furthermore, it appears that the data is significantly different from the binomial distribution (deviance and Hosmer-Lemeshow test revealed $p < .05$) therefore the distribution does not meet the conditions necessary to use this test. Since the data did not meet the condition of the logistic regression, a $\chi 2$ was performed on the above crosstab's table, but likewise did not yield statistically significant results $\chi 2$ (2, N = 54) = 4.162, p = .125.

The final spreadsheet project.

When first envisioned, the final modeling project was seen as the final instructional tool before the posttest assessment. However, it became apparent that this project provided a central assessment in its own right, as the only assessment of the students' ability to build a model from scratch (an assessment of modeling itself instead of merely an assessment of knowledge about models), which was lacking from the SUMS pretest and posttest.

The dissertation proposal gave a rubric that assesses students on the variable selection in the model, how these variables are integrated into the model, the level (concrete, formal, or post-formal) of the model, whether or not the model was checked against data, and the quality of the hypothesis that the student formed and/or tested with the model.

All aspects of the final modeling project were correlated to CTSR score. While not part of the original research proposal, the actual classroom grade on both the spreadsheet project ($r (58) = 0.402$, $p = 0.001$) and the paper explaining the project ($r (58) = 0.444$, $p < 0.001$) were

Table 13. *Preliminary variable list, relevant to irrelevant variable ratio results.*

CTSR	*Almost no relevant variables*	*More irrelevant than relevant*	*About equal relevant and irrelevant*	*More relevant than irrelevant.*	*Almost no irrelevant variables.*
1-8 (concrete)	2	1	1	3	1
9-14 (low formal)	5	1	1	6	6
15-20 (high formal)	1	0	3	7	10
21-24 (Post formal)	0	0	2	0	3

significantly correlated with CTSR total. In addition, the total of the rubric score (variable selection + variable integration + hypothesis testing + level of model + model checked against data) for each student was also correlated ($r(58) = 0.376$, $p = 0.004$,) with total CTSR score. Each of these correlations is moderate size.

Not only were the total scores for the project significantly correlated to the students' CTSR scores, but a binary logistic regression revealed that several of the sub-scores on the rubric were also related to CTSR scores, once the four point scale used to score these questions was collapse to a binary scale. The results of these analyses are presented in Table 14.

Table 14. Results of binary logistic regression of student project rubric sub-scores.

Sub-score	*p binary logistic regression*	*Scores*	*Number of students*	*CTSR mean*	*CTSR Standard deviation*
Variable	.029*	0	0		
Selection		1@	15	12.27	3.95
		2@	30	14.70	5.64
		3#	13	17.46	3.38.
Variable	.004**	0@	4	12.25	5.85
Integration		1@	19	12.11	4.36
		2#	26	16.19	5.28
		3#	9	16.67	2.96
Checked	.078	0@	16	14.94	5.48
model		1@	7	11.29	4.23
against data		2@	23	14.17	5.12
		3#	12	17.17	3.81
Hypothesis	.028*	0@	14	13.36	5.30
testing		1@	15	14.13	3.40
		2@	23	14.48	5.62
		3#	6	19.67	3.39

Table 14. Continued.

Sub-score	*p binary logistic regression*	*Scores*	*Number of students*	*CTSR mean*	*CTSR Standard deviation*
Level of model	.056	0@	3	11.67	6.81
		1@	10	12.20	6.01
		2#	45	15.40	4.59
		3	0		

Two students did not turn in a final modeling project. Both had low (9, 10) CTSR scores. If the assumption is made that they did not turn in the projects because they were not able to do the project successfully, and these points are scored as zeroes instead of being omitted from the analysis, then level of model (p = .026) also becomes significantly related to CTSR when analyzed by a binary logistic regression (in addition to the other three areas).

The first sub-score analyzed was *variable selection.* A score of *three* indicated almost no errors in *variable selection*, with all variables included being important, and no unimportant or incorrect variables included. A score of *two* allowed for some minor errors or incorrect variables. A score of *one* indicated major errors with *variable selection*. No student who completed a project was scored a *zero* on *variable selection*, as all students who submitted a project had at least some appropriate variables selected. Examples of appropriate and inappropriate/incorrect/ irrelevant variables are in Appendix E.

When the data in Table 14 are examined, there appears to be a clear relationship between the mean CTSR of students and their score on *variable selection*. Students scoring a *three* had

CTSR scores that averaged 2.61 points higher than students scoring a *two*, who had average scores approximately 2.43 points higher than students scoring a *one*. Since there was not enough data for an ordinal logistic regression, it was necessary to break this four level rubric into a binary rubric for analysis purposes. Unfortunately, where to draw this break (between a score of *one* and *two* or between a score of *two* and *three*) was not clear from the data. When the students with scores of *three* ($N = 13$, $M_{CTSR} = 17.46$, $SD_{CTSR} = 3.38$ were compared to the students with scores of *one* and *two* ($N = 45$, $M_{CTSR} = 13.84$, $SD_{CTSR} = 5.20$), a binary logistic regression did yield a significant result, $z = 2.18$, $p = .029$, and an odds ratio = 1.18. On the other hand, if the students with scores of *one* ($N = 15$, $M_{CTSR} = 12.53$, $SD_{CTSR} = 4.24$) are compared to the students with scores of *two* and *three* together($N = 43$, $M_{CTSR} = 15.40$, $SD_{CTSR} = 5.15$), a non-significant result is obtained from a binary logistic regression $z = 1.85$, $p = .064$, and an odds ratio = 1.12. However, if the two students not completing the assignment were scored as zeroes and included with the lower scoring group instead of being omitted, the result becomes significant ($p = .031$). In summary, the students receiving a score of *three* had significantly higher CTSR scores than the other students and there is support for the idea that the students scoring below *two* also had significantly lower CTSR scores than the other students.

Scoring for *variable integration* was similar to scoring for *variable selection*. A score of *three* indicated almost no errors in *variable integration*, with all appropriate variables present and connected by appropriate formula, and no unimportant or incorrect variables or relationships included. A score of *two* allowed for some incorrect or missing relationships. A score of *one* indicated major errors with the formulas/relationships between variables. Four students who completed a project were scored a *zero* on *variable integration*, as they had no formulas/relationships in their final project.

Both groups that had some success at *variable integration* (rubric scores of *two* or *three*) (N = 35, M_{CTSR} = 16.29, SD_{CTSR} = 4.74) had mean CTSR scores roughly four points higher than those students who had with little or no success (rubric scores of *zero* or *one*) (N = 23, M_{CTSR} = 12.17, SD_{CTSR} = 4.57) at *variable integration*. This data provided a much cleaner break for converting to a binary rubric for binary logistic regression than *variable selection*. Binary logistic regression did yield a significant result, with z = 2.85, p = .004, and an odds ratio = 1.20 (with p = .002 if the students not completing the assignment were scored as zeroes instead of omitted from the analysis).

The rubric score for *compare model to data* was on a scale of zero to three. A score of *three* indicates extensive comparisons with outside data/models to verify the correct behavior of the constructed model. A score of *two* indicated a comparison between models was made, and some analysis of how or why the results obtained by one model were the same as or different from the other model was made. A score of *one* indicated that students made a comparison to another model, but that comparison did not discuss how or why the results obtained by one model were the same as or different from the other model. A score of *zero* was given if no comparison to another model or outside data was made. As is discussed in more detail in Appendix E, there is compelling evidence that this lack of comparisons was not always due to purely student factors (i.e. availability of external models were not uniformly available).

While no clear pattern in CTSR scores appears when compared with rubric scores at levels zero, one and two, the CSTSR scores of students scoring a three (N = 12, M_{CTSR} =17.00, SD_{CTSR} = 3.81) on checking model against data are higher than the CSTR scores of the rest of the students (N = 46, M_{CTSR} = 14.04, SD_{CTSR} = 5.19). An analysis reveals that this relationship is not quite

statistically significant, with a binary logistic regression showing $z = 1.76$, $p = .078$, and an odds ratio = 1.14.

The rubric for *hypothesis testing* was again on a scale of *zero* to *three*, with a score of *zero* indicating that the student did not form a hypothesis at all. Despite its explicit mention in the directions, 14 of the 58 students turning in this assignment (or just under 25%) did not form a hypothesis at all, and received a score of *zero*. Another 15 wrote what they called a hypothesis, but the hypothesis did not appear to be based on the student's model in any way. These hypotheses received a score of *one.* Thus, exactly half of the students turning in the assignment did not use their models to form a hypothesis. By far the largest group of students (23 of 58) formed a trivial hypothesis based on the model and received a score of *two*. For the purpose of this study, a trivial hypothesis was defined as a mere extension of the original intent of the model. From the student directions:

> For example, if your model was paper versus plastic bags, how many pounds of CO_2 or units of energy would be saved by mandating a switch to using only the better bag? This type of hypothesis will be considered a trivial hypothesis because it follows directly from the model, if your output predicts that a paper bag saves $.03 over a plastic bag, then if 10,000,000,000 bags are used in the United States in a year, one only needs to multiply the above numbers to find a savings.

Only 10% of students formed a hypothesis that clearly demonstrated full formal reasoning, and received a score of *three*. The directions again specifically stated:

> A more interesting hypothesis would be to consider how changes in your input variables would affect the output (for instance, if your model was created three years ago with gas under $2.00/gallon, does the answer change if the price of gas goes up to

> \$3.30/gallon?) Another alternative would be to explore what value of a variable would be necessary to reverse your decision? What is the necessary price for a barrel of crude oil before plastic bags are the better option? At what price of landfill space does the option which produces the most garbage cease to be the cheapest option? What value must be assigned to a tree before the using of that tree as raw material becomes more expensive than leaving it in place to provide shade, provide CO_2 sequestration, prevent soil erosion, and other services?

Each of these paths represents another step in modeling and abstraction, to think about the input variables not in terms of what is, but in terms of what may be. Despite these instructions, only six students completed a hypothesis in which they predicted the effect of the change of at least one variable on the outcome of their model.

The *hypothesis testing* rubric score of the final modeling project showed little difference across the lower levels (students with rubric scores of *zero*, *one*, and *two* all had average CTSR scores within 1.12 points of each other). However, the six students with a rubric score of *three* ($N = 6$, $M_{CTSR} = 19.33$, $SD_{CTSR} = 3.67$) on *hypothesis testing* had significantly higher CTSR scores than the students receiving a score of less than *three* (i.e. *zero*, *one*, and *two* collapsed into a single group) ($N = 52$, $M_{CTSR} = 14.12$, $SD_{CTSR} = 4.94$), binary logistic regression $z = 2.20$, $p = .028$, and with an odds ratio = 1.33, which is quite high.

The *level of model* rubric score was also assessed on a scale from *zero* to *three*, with *zero* being a non-model (a table reporting static calculations, for example), *one* being a model with only concrete components (such as the tangible objects: miles, gallons, dollars, etc.), and *two* being a model with abstract or invisible components that should be familiar (such as molecules of carbon dioxide, the environmental cost of a tree, etc.). A level *three* model was not expected or observed,

but would have contained postulated components or combined components in a way that is outside the typical established relationships. Lawson (2002) and Lawson, Alkhoury, Benford, Clark, and Falconer (2000) describe a true scientific model such as a Mendel's gene model or Dalton's atomic model as such a model. The existence of an unknown, postulated structure with specific characteristics was necessary in each case, even though there was no direct evidence that such objects existed.

A binary logistic regression was performed comparing the *level of model* to the CTSR total score. The student models scored as *zero* or *one* (both not containing abstract components) were collapsed into a single category (N = 13, M_{CTSR} = 12.23, SD_{CTSR} = 6.10) and compared against the student models scoring *two* (which did contain abstract components) (N = 45, M_{CTSR} = 15.36, SD_{CTSR} = 4.55). The result was not significant when analyzed with a binary logistic regression, z = 1.91, p = .056, and an odds ratio = 1.14. If the 2 students not completing the project were scored as a zero instead of omitted from the analysis, the p-value drops to .026, which would be significant.

In conclusion, the final modeling project provides strong support to the idea that cognitive developmental level played a large role in a student's success at constructing a model. Not only was almost every rubric sub-score able to be significantly related to cognitive development, the other measures of the modeling (classroom grades on the model and the paper accompanying it) were also significantly correlated with CTSR scores.

Threats to Validity

There were several potential threats to the validity of this study. Some concern the help that students were able to access that might have caused their scores on modeling projects and activities to be higher than they should. Others concern difficulties with scoring and the sample.

Inaccurate representation of student ability.

The multiple roles of the instructor/researcher provided an opportunity for several threats which will be discussed first. Although the instructor/researcher is relatively common in this field, including many of the studies cited previously (Akerson, V. L., Abd-El-Khalick, F. & Lederman, N. G. (2000);. Sarri, H. & Viiri, J. (2003); Schwarz, C., & White, B. (2005); Windschitl, M. & Thompson, J. (2006)) this dual role poses conflicting agendas at times that should be responsibly examined.

Nowhere is this potential for conflict more obvious than in the central goal of each role. As teacher, the goal is for every student to succeed. As researcher, the goal is to discover if low CTSR scores contribute to student failure in modeling tasks – thus failure is needed, and failure by students with low CTSR scores would help support the hypothesis. On the other hand, more apt students may be more enjoyable to teach, and giving them specific help (consciously, unconsciously) could lead to greater gains by students with high CTSR scores which would also help to support the hypothesis. Examination of the videorecordings of the lessons as well as email records show evidence that neither of these occurred.

Email records from the time of the final project show more numerous interactions with many of the students with low CTSR scores who were struggling with the final modeling project than with students with high CTSR scores. For example, the two students receiving the most help via email (and via office hour visits) both had low CTSR scores (<10) and received 11 emails (in a two week period) and 16 emails leading up to the submission of the final project, far more than were exchanged with any other students. Furthermore, this email quote from the student receiving 11 emails reflects the extent to which extensive help was given to aid students (all typographical errors present in the original email) “hey i know you already helped me SO much, but you gave

me list to but [put?]in my excel sheet, like pollution, pollution cost you listed others and when you left i was trying to write them down, but i have a horrible memory and forgot what all you said. I'm SO sorry. I just feel totally lost. Thanks for everything ." Additionally, this quote captures the essence of how the dual roles were handled. During office hours, ways to transform her draft spreadsheet into a more appropriate spreadsheet were discussed, including trying to get her variables for various pollutions to a single unit (dollars of environmental damage). Intentionally, this discussion was carried out in such a way that the student was not able to bring written records out of the conversation of exactly what variables to include, nor were these provided via email. Review of recordings supports this claim. The intention was that if these concepts were within the student's zone of proximal development, the student would be able to use this nudge to make appropriate changes to the model. If, as appeared to be the case with this student, this concept was so far above the student's understanding that the student was unable to retrace the thinking once they left the instructor's presence, the final model might not improve. Students were able to repeat the discussions, but the "list" of variables that the above student requested were never directly provided in a format the students might be able to use without understanding what they were doing.

Therefore, the researcher, in his role as teacher, may have influenced the results of the student success. Students with lower CTSR scores, but who were perhaps on the cusp of understanding how to create models, received help that might have allowed them to make connections and complete a better model than they might have otherwise constructed. However, since this influence would tend to diminish a correlation between student success and CTSR score, it is likely that the results on the final modeling project (and to a lesser extent, the small modeling

assignments) may underrepresent the importance of cognitive ability through the instructor/researcher's attempts to help all students succeed.

A second potential threat also relates to unrepresentative scores on the final project and other modeling assignments due to the influence of other classmates. While the assignment was designed to be an individual assignment, it was apparent that some amount of collaboration occurred. Several pairs of students worked on similar final projects. In some cases, these friends or couples had drastically different CTSR scores. Their final modeling projects (and to a lesser extent, small modeling assignments) were similar and at the level corresponding to the student with higher ability. Like the assistance from the instructor discussed above, this threat would tend to lessen, not strengthen, the relationships between CTSR score and student success observed, thus underrepresenting the importance of cognitive ability.

Scoring issues.

The scoring of small modeling assignments, the final project, and free response questions on pretest and posttest was subjective. Use of an additional scorer for pretest and posttest helped, although this was not as effective as desired. In several cases, the second scorer misapplied the rubric. For example, the second scorer tended to interpret the word *accurate* differently than intended. While the word *accurate* was part of the rubric for the pretest/posttest question 13 on multiple models (where discussion of the *accuracy*/reliability tradeoff indicated a score of three) and question 14 on the most important aspect of a scientific model (where the ability to make *accurate* predictions indicated a score of a three) the second scorer tended to give a score of three whenever the word *accurate* was used, regardless of context. In some cases, the context clearly indicated that the student meant that the model accurately depicted in the physical object, which was more consistent with a score of one. Furthermore, there were two instances where the single

word *accurate* was given as a free response answer with no context. Again, the second scorer tended to score these higher (three points) than the researcher (one point). Despite these issues, initial inter-rater reliability (Spearman's Rho) varied widely, but was higher for the relatively straightforward scoring question 12 (types of models, $\rho = .694$), question 11 (theory change, $\rho = .813$) and question 13 (multiple models, $\rho = .634$) to the much more difficult scoring question 6 (hypotheses, theories and laws, $\rho = 0.214$). Question 15 (model construction, $\rho = .300$) and question 14 (most important aspect of a model, $\rho = .549$) fell in between.

While inter-rater reliability is one method of determining the reliability of scores, four of the sub-scores offered the opportunity for triangulation between Liker-scale and free-response portions of pretest and posttest. Pearson correlations for each of the four free response questions where this was possible were all at least moderate: question six ($r = .346$), question 11 ($r = .458$), question 13 ($r = .378$), and questions 14 ($r = .455$). Thus, while inter-rater reliability was low for question six, the recorded scores were consistent with the student Liker-scale responses, which increased faith in these scores.

A few questions were difficult to score because of their wording. Some questions had too many parts, and should have been broken down into separate questions. On the pretest/posttest, the most difficult question was question six, which asked students for three definitions and up to three pairwise comparisons. With so many required parts to the answer, it was not clear when students answered only part of the question if they omitted the rest of the answer because they did not know the answer or because they forgot that there was a second part to the question. This multipart complication was also present in Carbon Footprint question two.

Another way that the these rubrics were problematic was in showing small movements in performance, because two answers could receive the same score but show slight but definite

differences in understanding. Rescoring pretest and posttest side by side gave a finer view for changes that might not represent a whole point move. Scoring side by side allowed the researcher to examine if there was any difference in a *two* on the pretest from a *two* on the posttest. This could represent an answer better in one subpart but worse in another. In at least question six, since there were more subparts than points, a gain in a subpart might not be enough to merit a gain of a whole point on the rubric, however a definite change was present. Appendix E does discuss that actual gain, if these half steps were included might have been larger still.

A final threat was significantly less detailed posttest free response answers for some students vs. their pretest free response answer to the same question. For example, there were students who gave extensive answers that earned points on the pretest, but these same students left these same questions blank on the posttest. While this apparent lack of effort on the posttest was not widespread (two students) it was as common as not for posttest free response answers to contain fewer words than their corresponding pretest answer. Normalized change (as opposed to normalized gain) was used to minimize the negative impact of posttest scores that were lower than pretest scores.

CHAPTER FIVE

CONCLUSION

Perspective for the conclusion

The literature review described a number of studies at a variety of levels that have attempted to measure and understand student knowledge of modeling and ability to construct models. The literature review also described research into student understanding of the nature of science (NOS), instruments for measuring NOS and strategies for improving student understanding of NOS. Piagetian development also has well-developed instruments, protocols and body of knowledge described in the literature review. This study has attempted to establish a relationship between these three areas, an area of overlap demonstrated to be lacking from the literature. This study attempted to provide insight into whether or not a student's Piagetian developmental level influences how much a curriculum designed around several incrementally more complex modeling activities results in deeper understanding of models and the nature of science, specifically: (a) the relationship between theories, laws, models, and hypothesis; (b) how and why theories change over time; (c) how and why models are refined; (d) the purposive nature of model creation; and (e) the role of models in scientific investigations.

Research question and sub-questions.

Is attainment of the formal operational Piagetian level of understanding necessary for a model-based environmental science curriculum to increase students' understanding of models and the nature of science?

Sub-questions:

1. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations), and is that improvement related to Piagetian level?

2. Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the utility of models (communication, simplification for study, prediction), and is that improvement related to Piagetian level?

3. Does a curriculum emphasizing student comparison, refinement, and creation of models improve student understanding of the relationship between models, theories, and the scientific method (models operationalize theories, allowing them to be tested with the scientific method), and is that improvement related to Piagetian level?

Hypotheses.

The null hypotheses were that there will be no significant or important difference at the p = .05 level in student understanding of models nor understanding of the nature of science before and after completing a semester of the model-laden environmental science curriculum. There will be no significant difference at the p = .05 level between any normalized gain between the pretest and

posttest in modeling and/or nature of science knowledge between students in the post-formal operational stage, formal operational stage, early operational stage, and pre-operational/concrete stage of cognitive development. (This curriculum includes exposure to authentic model use, critique and modification of existing models, comparison of multiple models of the same system, analysis of the conscious choices that shape models, and construction of models and use of these models to answer questions.)

The alternative hypotheses are that there will be statistically significant gain in students' modeling knowledge and/or nature of science scores on the posttest as compared to the pretest. This difference will also have normalized gain of greater than 0.5 (medium effect). Furthermore, when any gains in modeling and/or nature of science knowledge are correlated to the cognitive development of the same student, it is expected that students who have reached a higher operational level of development (post-formal > formal > transitional > pre-formal) will have statistically greater gains than students with lower levels of development.

Data summary.

Chapter four provided detailed examination of the quantitative data collected during the study. The Classroom Test of Scientific Reasoning (CTSR) measured students' cognitive development and established the independent variable. Individual student modeling assignments provided both quantitative data and qualitative information regarding these questions, particularly understanding of models and modeling. The final modeling project provided both quantitative data and qualitative information regarding students' understanding of models and modeling and also determined whether or not a student could actually apply this knowledge towards building an

actual model. The SUMS/SUSSI pretest and posttest taken together measured the gain in student knowledge regarding models and modeling, as well as several nature of science concepts.

The purpose of these next sections is to cross-reference the results of the individual sources of information against the research question and sub-questions. For each section, improvement will be examined first (pretest-to-posttest gain), followed by information related to this area gathered during the individual assignments and modeling project. Finally, the correlation between these gains and CTSR score will be discussed along with correlations between student success on a task and CTSR score.

Research Sub-question One: Nature of Models

The stated question from the proposal was: *Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations), and is that improvement related to Piagetian level?* The SUMS categories related to this question would include models as *exact replicas* and *multiple models.* In addition to the existing SUMS categories, question 12 on the pretest and posttest measured student's knowledge of the *types of models*, which seems to be related. In addition to the pretest and posttest, the Resource Use Activity and the Carbon Footprint Activity were particularly rich sources of data for this question.

Models as exact replicas.

Questions 16, 19, 33, 34, 35, 36, 37, and 38 on the pretest and posttest were used to assess students' knowledge and misconceptions about models as *exact replicas*. Several of these questions (16, 19, and possibly 35) were revealed to be confusing in the follow-up interviews (see Appendix E). Only question 36 dealing with the face validity of models showed any real gain

(Cohen's $d = 0.71$, normalized change 0.34). Overall, this sub-score showed among the lowest gains related to modeling, with an Cohen's $d = 0.42$ (corresponding to an average raw gain of 0.36 points out of a possible eight) as students moved from a pretest class average of 4.68 to a posttest class average of 5.04, a normalized change of 0.21. Therefore, by all accounts, scores in this sub-score showed at best small gains and were at worst largely unchanged.

Furthermore, the gains that were shown were not as correlated to CTSR score as several of the other sub-scores were. Overall, the correlation between residual gains in this sub-score and CTSR score of students was $r = 0.29$, a small-moderate correlation. This correlation, however, was significant ($p = 0.027$).

Finally, it should be noted that none of the specific classroom activities addressed the bulk of these questions – the research design wrongly assumed that students would make gains in this area without the explicit reflection on these topics. The large (and unique for this sub-score) gain on question 36 makes sense when one considers that students did critique models in each of the classroom activities (Human Population Lab, Carbon Footprint Activity, Resource Lab, and Global Warming Model), which logically could instill in students a need for a model to have valid inputs. Students by and large were very successful in this critique. In the Human Population Lab students who wrote a successful critique had CTSR scores an average of three points higher than those who did not, although the logistic regression was not significant ($p = 0.051$). This result is close enough to significance that it would be worth repeating the experiment with a larger sample size. The Resource Lab asked several questions relating to this area. Question one, which asked which variable to delete did not find broad success, with only 14 students (of 53 answering the question) who answered this question correctly. However the students answering correctly had mean CTSR scores nearly six points higher than their unsuccessful counterparts, and a binary logistic

regression revealed that CTSR was significantly related to this success ($p = 0.024$). Question two asked students what variable to add, with broader success (42 of 53 answering were successful) and better significance related to CTSR score ($p = 0.014$) as determined by binary logistic regression. Question three, regarding what variable to change also had broad success (43 of 53 answering the question were successful) and significance related to CTSR score ($p = 0.032$) on binary logistic regression.

The Carbon Footprint Activity also asked students to critique and compare the various inputs in question two (several different models in this case). Again, most students (39 of 52) were completely successful, with a further nine having partial success. Average CTSR scores for those who had at least some success were three points higher than for the four students who did not, although this was not statistically significant, as would be expected by such an unbalanced sample. The Global Warming Activity, question eight, also related to the quality of the inputs of the model in question. Although this question also included some hypothesis testing aspects, 23 students answered this question correctly, and this group had significantly different ($p = 0.025$) CTSR scores from the CTSR of students who answered the question incorrectly. Thus, with all of these opportunities to critique models, it is not surprising that students showed gains on this single question.

On the other hand, it was expected that by showing mathematical models of systems and other non-physical phenomena which in no way looked like the phenomena they were modeling, that students would have stronger gains in the models as *exact replicas* categories. This did not occur. In hindsight, it would have been very helpful to have put questions in some of the earlier activities explicitly asking students to reflect on whether or not the model that they had just used or made physically resembled the phenomena or was an exact replica in every way except for size.

Perhaps without leading those students explicitly to those conclusions in the activities themselves, it was not possible for students to overcome these deeply held misconceptions of models as primarily physical models. Additionally, the ways that models can be used at the second and third level do not preclude their use at lower levels when appropriate – a level three modeler can use a model to see structure or instruct another.

Therefore, in conclusion, there are two separate conclusions for the sub-score *exact replicas*. For the question of evaluating the inputs of a model for face validity, scores from the pretest to posttest show a strong effect size. This gain can be explained in terms of the classroom activities, where most students were able to find success. In addition, this success was often linked to CTSR. For the sub-score as a whole, there was little gain, as the face validity question was diluted over a number of other questions relating to aspects of physical models. These issues were not explicitly dealt with in class. Therefore, it may be necessary to explicitly ask students to reflect on these issues of models as exact replicas if gains in this area are desired.

Multiple models.

Questions 13, 26, 27, 28, 29, 30, 31 and 32 on the pretest and posttest measured student understanding of multiple models. Questions 13 (free response), 28, 31, and 32 all showed large (0.7 or greater) Cohen's *d* statistics. Only question 26, which had serious wording concerns, failed to show gain. The sub-score as a whole showed an average pretest score of 4.85 (out of 10), an average posttest score of 6.32, a raw gain of 1.46, a normalized gain of 0.39 and an Cohen's d = of 1.40. Specifically, free response question 13 showed increase in the number of students discussing how multiple models (a) reflect different choices in complexity and accuracy, (b) are made for different purposes and (c) reflect different aspects of the same phenomenon. Conversely, in question 13 there was a decrease in the number of students who wrote answers consistent with the

misconception that two models of the same object could exist only if at least one was incorrect. Thus, by almost every measure, students showed strong gain on the sub-score of *multiple models*.

In addition to a strong gain in understanding demonstrated from the pretest to the posttest, the gain in *multiple models* showed correlations with CTSR score. A regression of *multiple models* sub-scores residual gains on CTSR showed a correlation of $r = 0.37$ (moderate). In addition to the strength of the correlation being moderate, this correlation was also significant ($p = 0.004$). Therefore, gains in the *multiple models* sub-score were not only large and significant, but also significantly related to CTSR score.

Finally, two of the activities measured student understanding of multiple models. The first activity to focus on multiple models was the carbon footprint activity. Each of its three questions focused on some aspect of multiple models. Question one showed 43 students with some success discussing how multiple models could arrive at different conclusions and how these differences could be interpreted appropriately, while nine students did not find success, and eight did not answer. Those students answering the question at least partially correctly had mean CTSR approximately four points higher than those who answered incorrectly, and this difference was significant when tested with a binary logistic regression ($p = 0.018$). Question two was summarized previously as it also related to the *exact replicas* sub-score, but briefly 48 students had success, four students did not, the CTSR scores of those who had success were larger, but no statistical significance could be shown in part because the number of students not having success was so small. Question three showed 47 students were able to correctly state explicitly how two of these multiple models could have been used differently while only seven students were not able to successfully answer this question. The mean CTSR of the students who could was almost 2.5 points higher than the mean CTSR of the students who could not, although CTSR was not

significantly ($p = 0.201$) related to the score on question three when a binary logistic regression analysis was performed.

Finally, question seven of the Global Warming Activity asked students to compare the results of multiple models and interpret the meaning of the disagreement in exact conclusions reached by the eight models. Forty-three students agreed that these differences did not make the multiple models invalid or inaccurate, and only 12 students who answered did not agree. Students who answered correctly had mean CTSR scores higher than those who did not, with a binary logistic regression analysis showing this result was significant ($p = 0.019$).

Thus, in conclusion, students showed success during the activities throughout the semester relating to multiple models. This success translated to large gains on the posttest. Both the success during the activity and the gains were typically shown to be significantly related to CTSR score.

Types of models.

This sub-score was represented by a single question (question 12) on the pretest and posttest, and not explicitly addressed in the individual activities. As stated in the data analysis, large gain in this sub-score was seen, with a mean pretest score of 0.87, a mean posttest score of 1.77, a raw gain of 0.90, and a Cohen's $d = 1.15$ and normalized change of 0.39. Most, but not all of this gain can be seen in a large increase in the number of students who listed mathematical models in the posttest, as would be expected with the level of student involvement with mathematical models in this class. In addition, however, 13 students added two additional types of models on their posttest as compared to the pretest, thus these gains were not entirely a result of working with the mathematical models.

These gains were not, however, well correlated with CTSR score. Perhaps because a) gains were so universal, b) there were only four levels of this variable in a single question, c) essentially this task asked for a regurgitation of facts, and d) the learning was so directly linked to classroom experience, the correlation between CTSR and residualized gain in *types of models* was small (r = .1631), and not significant (p = 0.213).

Although the activities themselves contained no explicit reflective questions regarding types of models, the Global Warming Activity contained mathematical, visual, and conceptual models, and each of the other activities was a mathematical model.

Conclusion.

With regard to research sub-question one, significant evidence was collected demonstrating gains in several areas related to *understanding of the nature of models (model as a representation, multiple models, appropriate application and limitations).* The strongest gains and correlations were found with respect to *multiple models*, with strong gains also found in *types of models*. These gains largely exceeded the *a priori* Cohen's d = of medium (0.5), and the correlation of the gain to CTSR score, in the case of multiple models, a binary logistic regression received a medium correlation (0.37) as well. Only in models as an *exact replica* did the sub-score show little gain, nor correlation to CTSR.

Research Sub-question Two: Utility of Models

The stated question from the proposal was: *Does a curriculum emphasizing student comparison, refinement, and creation of models improve understanding of the utility of models (communication, simplification for study, prediction), and is that improvement related to Piagetian level?* The SUMS categories related to this question would include *uses/purposes of scientific*

models, *changing nature of models*, models as *explanatory tools*. In addition to the existing SUMS categories, an additional category that emerged was *how are models created*? In addition to the pretest and the posttest, each of the small classroom assignments (except for the human population lab) asked students to form a hypothesis and comment on the complexity/accuracy tradeoff, and thus can provide some data. The final modeling project provides the best data related to the category *how are models created*.

Uses/purposes of scientific model.

Questions 13, 14, 20, 21, 22, 28 and 29 from the pretest and posttest related to *uses and purposes of scientific models.* Questions 13 and 14 were free response questions; question 13 was cleanly scored, question 14 presented some difficulties (see Appendix E). Question 13, as discussed previously in multiple models, showed large gains in effect size (Cohen's d = 1.24) and normalized change (0.51) with 28 (of 60) students improving by one point (of three) and 11 improving by two points. Question 14's gains were likewise large, with Cohen's d = 0.72 and an average normalized change of 0.35, with 15 students (of 60) showing gains of one point (of three) and 15 students showing gains of two points. The most common reason students scores on the posttest were better than on the pretest was the indication that the student understood that scientists use models to make predictions and form hypotheses. The Likert-scale question 20, 22, and 28 showed strong gains and little confusion, questions 21 and 29 were slightly less clean and showed somewhat less gain. As a whole, the sub-score showed an increase from an average of 5.55 (of 11) on the pretest to 7.73 on the posttest, resulting in a normalized change of 0.39 and a large Cohen's d = 1.32. Therefore, it can be concluded that students showed large gains in understanding of *uses/purposes of scientific models*.

Not only were these gains large, but they were correlated with CTSR score. A moderate correlation ($r = 0.3838$) was achieved by regressing residual gains in the uses/purposes of scientific models sub-score on CTSR, and this correlation was significant ($p = 0.028$). Therefore, not only were these gains large, but the gains were statistically related to CTSR.

Finally, with regard to classroom activities, data collected on student use of language pertaining to *uses/purposes of scientific model* supports knowledge of where in the curriculum these gains may have started to occur. The Carbon Footprint Activity question one, asked about model uses in general as discussed previously in the section on *multiple models*. Briefly those students answering the question at least somewhat correctly had mean CTSR approximately four points higher than those who answered incorrectly, and this difference was significant when tested with a binary logistic regression ($p = 0.018$). Question three of the Carbon Footprint Activity asked students to discuss how two of the models examined could be used, and this question was also discussed in more detail previously in *multiple models*. Briefly, the CTSR of the students who could was almost 2.5 points higher than the mean CTSR of the students who could not, but CTSR was not significantly ($p = 0.201$) related to the score on question three when a binary logistic regression analysis was performed. Thus, overall, being able to understanding how models are used shows some signs of being significantly associated with formal thinking. In addition to these explicit questions, eight students during Resource Lab commented on the purpose or bias of the model, whereas in the following activity (Carbon Footprint), 15 students specifically mentioned the purpose or bias of the modeler. While two data points is very weak, a tentative trend of increasing knowledge about the purpose of models is established. While these questions asked about use of models in general, the most significant purpose of a scientific model is it to form accurate hypotheses.

Specifically, almost every activity asked students to make a hypothesis with the model. Question four of the Resource Lab asked students to form a hypothesis with the model. Of the 53 students attempting this question, 28 formed an acceptable hypothesis, and the mean CTSR score of the students forming an acceptable hypothesis was almost three units higher than the mean CTSR from those not forming a completely acceptable hypothesis, and this difference in CTSR scores of those who were able to form hypotheses and those who were not able to form hypotheses was significant ($p = 0.028$) when analyzed with a Binary logistic regression. Question eight of the Global Warming Activity also asked students to form a hypothesis, and again, students who successfully formed a hypothesis ($n = 23$) had CTSR scores approximately three units higher than students who did not provide a good hypothesis, and this result was significant ($p = 0.025$).

Students were also supposed to form a hypothesis in their final project. While half (29 of 58) of the students were able to form a hypothesis using their model, only six were able to form a hypothesis regarding how the change in one variable could affect the output of the model. These six students had a significantly ($p = 0.028$, binary logistic regression) higher CTSR scores than the students who did not form such a hypothesis.

In addition to activities where forming a hypothesis was explicitly stated, 15 (of 56) students made a hypothesis or otherwise reasoned using the model in the carbon footprint activity. Although the mean CTSR score of those reasoning with the model was higher than the mean CTSR score for those who did not, it was not significantly so.

In conclusion students showed moderate gains from pretest to posttest in the category of *uses/purposes of scientific models*. Not only were these gains meaningful in size, the gains were correlated to CTSR score. Statistics from the classroom activities support student success with

these tasks throughout the semester for most students, and this success was often related to CTSR score. Thus, this sub-score appeared to support research sub-question two.

Changing nature of models.

Pretest and posttest questions 23, 24, and 25 measured student understanding of the changing nature of models. Each of these Likert-scale questions scored cleanly with no misunderstandings. Pre-test scores were high (2.45 out of 3.00), potentially providing somewhat of a ceiling effect, as the posttest showed only a small raw gain of 0.23 to 2.68. Despite this, other measures of gain were still fair, with a 0.37 normalized change and a Cohen's $d = 0.64$.

On the other hand, the residual gains were not correlated to CTSR scores. There are several possible reasons for this lack of gain correlation other than there is no relationship. First, the category is small, with only three questions and three possible points. There is less spread available with this score as was seen with the types of models sub-score above. Second, the pre-test score was very high, providing a ceiling effect. One support of this idea is that both the pretest and posttest showed some correlation with CTSR: pretest $r = 0.197$ and $p = 0.131$ and posttest $r = 0.20$ and $p = 0.123$. Thus, with a high pretest score with little chance for gain already correlated with CTSR, there just is not adequate room for those students with high CTSR scores to show proportionally more gain.

While the pretest and posttest measured the *changing nature of models* somewhat less thoroughly than other categories, this category was well assessed in the assignments. Most of these questions have already been discussed previously, when discussing the gains seen in question 36 in the section on *Models as Exact Replicas.* These findings are summarized in table 15.

Thus, these data support generous opportunity to practice critiquing and changing models, overall good success in doing so, and often this success was significant at the 0.05 level by binary

Table 15. Summary of changing nature of models questions, described elsewhere in chapter five.

Activity	*CTSR of students who were successful vs. unsuccessful*	*Significance of relationship of success to CTSR (p) Binary logistic regression*
Human Population Lab	40 fully successful, CTSR 14.80 15 unsuccessful, CTSR 11.93	0.051 (not significant)
Resource Lab, Q1	6 fully successful, CTSR 19.50 47 unsuccessful, CTSR 13.92	0.024* (significant)
Resource Lab, Q2	42 successful, CTSR 15.12 11 unsuccessful, CTSR 10.63	0.014* (significant)
Resource Lab, Q3	42 successful, CTSR 14.93 10 unsuccessful, CTSR 10.90	0.032* (significant)
Carbon Footprint, Q2	48 successful, CTSR 14.24 4 unsuccessful, CTSR 11.50	0.264 (not significant)
Global Warming, Q8	23 successful, CTSR 16.39 27 unsuccessful, CTSR 13.30	0.025* (significant)

logistic regression, establishing a relationship between success on this task and CTSR score. Finally, the final project itself was submitted in various drafts. Although no formal data was collected on the drafting process that would fit here, this process should have reinforced the idea

that models could be changed. In conclusion, this category showed good student understanding on the pretest that was already somewhat correlated with CTSR. Through the activities, students had ample opportunity to critique and change models, and generally over 2/3 of students were successful (and in cases where they were not, the question involved another part that made it more cognitively demanding, such as Q8 on Global Warming which also involved hypothesis testing). This success appeared to be significantly related to CTSR score. Thus, although on posttest, overall gains were moderate (due to ceiling effect) and correlations to CTSR were weaker (due to broad success among both students with high and low CTSR scores) there appears to be ample support through the classroom activities that cognitive development is related to understanding the changing nature of models.

Models as explanatory tools.

Questions 16, 17, 18, 21 and 28 on the pretest and posttest measured the category models as *explanatory tools*. Each of these was a one point Likert-scale question giving a total of five possible points in this category. Question 16 had multiple words that caused confusion to students. Questions 17, 18, 21, 28 were much less problematic, although a few issues (definition of *phenomenon,* confusion about what a scientist actually does?) came up in follow up interviews. The pretest showed a category score of 3.14, with a raw gain of 0.37 to 3.51 on the posttest. This gain was moderate with Cohen's $d = 0.57$ and normalized change (0.22).

While gains were moderate, they were not particularly related to CTSR score. The correlation between CTSR score and residualized gain from pretest to posttest was small-moderate ($r = 0.25$) and this correlation was not quite significant ($p = 0.055$).

Virtually no individual questions directly asked students to reflect only on *models as explanatory tools*. Some students in the Human Population Lab commented on whether or not the

HDI was a good model for predicting or explaining the quality of life in a country, and supported their conclusion with examples where the model explained the quality of life well or poorly, but the idea of using a model as an explanatory tool was not explicitly asked in the question. Likewise, some students, in question three of the Carbon Footprint Activity discussed how some models were more suited to explain the idea of a carbon footprint to younger students and some models were more suited to a more sophisticated audience. Finally, in the first part of the Global Warming Activity, students used an explanatory model/simulation to attempt to understand how the Greenhouse effect works. Again, however, these questions were not explicit reflective question about that model as an explanatory tool but rather questions focused on the student's understanding the scientific concept attempting to be explained by the model. In hindsight, this sub-score, much like models as *exact replicas*, suffered from the methodological flaw that these ideas were seen as so simple (as opposed to hypotheses, etc.) that gains would occur without explicit reflective questions. Since the primary goal of this study was to move students to the highest level of scientific model understanding, reasoning with models and using them to make hypotheses, less explicit attention was paid to level two (models as explanatory/teaching tools) goals.

Question 21 deserves further discussion. It states "Scientific models' primary value is in showing/teaching science." Students *agreeing* with this answer are demonstrating a level two understanding of models, which is better than disagreeing because the student is stuck in level one and considers models an exact physical replica. On the other hand, students agreeing with this statement have not progressed onto the third level. The desired response is *strongly disagree* (since at the highest level models are used by scientist to make predictions about the behavior the system being modeled). During the follow-up interviews with students to discuss their pretest and

posttest answers, students showed no evidence that they understood how models could be used by scientists except perhaps to explain things to each other. Therefore, too much emphasis on models as teaching tools, without explicit examples indicating level three model use, runs the risk of halting student growth at level two.

In conclusion, while the data shows that gains did occur in student understanding of *models as explanatory tools,* and while the correlation showed that there might be merit in further pursuing the relationship between cognitive development and gain in this category, the most obvious conclusion is that this category of understanding models would have been best served by a stronger methodology with explicit reflection.

How are models created?

This category, like *types of models,* was also not part of the original SUMS and its construction more closely followed the initial methodology presented in Grosslight, Unger, Jay, and Smith (1991). In their study, Grosslight et al. directly asked students to name as many kinds of models as they could. Treagust et al. (2002) excluded this aspect in their study. Furthermore, this category was not described in this research proposal (based on the development of the SUMS) and instead emerged from an analysis of the data, which seemed to suggest that although some students seemed capable of knowing *about* models, they were not capable of the process of modeling. Therefore, a way to measure gain was needed. A test is not the appropriate way to measure performance of a process, and thus information for this section comes primarily from the final modeling project. Nonetheless, question 15 (three point, free-response) was a natural fit for this category and question 36 (Likert-scale) was also deemed to assess a basic component of building a model. This gave a total of four points for this category. The average pretest for this

category showed 1.99, with a raw average gain of 0.89, yielding an average posttest for this category of 2.88. This gain resulted in a normalized change of 43% and a large Cohen's $d = 1.03$.

Perhaps because of the very large gains seen across the board, there was no correlation seen between CTSR score and normalized change in this category. If anything, the treatment appeared to lessen a correlation that existed when the students entered the class, as the pretest scores of *how are models created* were significantly correlated with CTSR ($p = 0.031$) but the posttest was not ($p = 0.137$).

Obviously, however, the ultimate measure of student's ability to understand a process goal is to ask students to carry out the process, rather than merely asking students about the process. Thus, the final project becomes the primary source of data for this sub-question. The preliminary variable list assignment saw differences in the quality of variables listed that seemed to be related to CTSR scores, however, no statistical relationship could be found as the binary logistic regression was not able to be used due to the five categories of variables and a chi-square was not significant $\chi^2(2, N = 53) = 4.162$, $p = .125$. With regard to the actual project, however, students with higher CTSR scores were significantly more successful at selecting variables ($p = 0.029$), integrating variables ($p = 0.004$), and making hypotheses with the model ($p = 0.028$), with all tests being binary logistic regressions. In addition, most other measures of the final modeling project were close to significant, such as checked model against data ($p = 0.078$) and the level of the model itself ($p = 0.056$), again using binary logistic regressions. Thus, it can be concluded from the final modeling project that the ability to create a model (at least a mathematical spreadsheet model) seems to be related to cognitive development.

Why are the results of modeling and understanding models so different? The most straightforward answer is that questions 15 and 36 are asking for lower level cognitive processes.

Question 36 asks students to recognize a fact: that good models are largely made up of variables that make sense logically when the phenomenon in question is considered. Question 15 is perhaps slightly higher cognitively, asking students to recognize which of the models examined in class the model in the question most closely resembled, and then apply knowledge about how the class model was created to explain how the model in question 15 might have been created. Creation of a model, on the other hand, is at the highest cognitive level. It is at this higher cognitive task level that developmental level seems to be a more determining factor for success.

An everyday analogy might help to illustrate this concept further. Interested spectators of many performance events, whether it is athletics or arts, can describe the basic fundamental concepts of event. A person can recognize the use of light, color, texture, etc. in a masterpiece without being capable of creating a masterpiece themselves. Likewise a person can understand meter, harmony, melody, chord progressions, even recognizing more specific techniques such as suspended chords resolving to a major, without being able to perform the pieces themselves, let alone create them. Regardless, it would seem that some students performed like the master and other students like a knowledgeable audience, and more likely than not, the students who created the model had a higher CTSR score than those who could merely appreciate the general ideas of building a model when they observed one.

Conclusion.

As with the first sub-question, the data supporting the second sub-question is mixed. Students definitely showed large gains, with the smallest gains still showing moderate Cohen's d = 0.57, and two categories showing gains over 1.00. Thus, the first part of the sub-question, that gains will be shown, is clearly met. The second part of research sub-question two, however, is that these gains are related to developmental level as measured by CTSR score. The results showing

this relationship are somewhat less clear, with only uses/purposes of scientific models showing a clear statistical correlation of pretest/posttest normalized change with CTSR score. Other categories tended to have *p*'s close to, but not within statistical significance. Finally, regarding the ability to actually make models, developmental level seemed to be quite well related to student success in the final modeling project. Although this ability was not measured pretest/posttest and so gain could not be determined, students who were of high developmental level appeared to leave with a good understanding of how to build and use a model, regardless of where they started with this understanding.

Research Sub-question Three: Nature of Science

The stated question from the proposal was: *Does a curriculum emphasizing student comparison, refinement, and creation of models improve student understanding of the relationship between models, theories, and the scientific method (models operationalize theories, allowing them to be tested with the scientific method), and is that improvement related to Piagetian level?* There were three areas of the pretest and posttest supporting this sub-question: *nature of hypotheses, theories, and laws; theory change; and scientific method*. The information on this sub-question comes almost exclusively from the pretest and posttest, as this area was not taught directly in class, and students were not asked to reflect on these ideas. One exception would be that students formed model-based hypotheses repeatedly throughout the course, but this connection was not made explicit during the activities so the students were not able to comment on these connections. Since the connection between NOS and models was strongest through theories, and weaker in other areas, it was expected that students would show the most gains in theory change, and less gain in the other areas.

Theory change.

Questions 7-11 on the pretest and posttest measured students' understanding of the process of theory change. These four Likert-scale (one point each) and one free response question (three points) showed strong pretest results (5.18/7.00) but only modest raw gain (0.31) for a posttest of only 5.49/7.00. The results of these small raw gains were not surprisingly small Cohen's $d = 0.32$ and normalized change of 0.24. These results are quite a large decrease from the Cohen's $d = 0.64$ and normalized change of 0.37 seen for the *changing nature of models*. Since the logic chain was that students would use actual experience changing models to learn about the *changing nature of models*, and by extension, theories, it would be expected that gains in *theory change* might be somewhat smaller but on the same order as the *changing nature of models*. Again, a stronger explicit reflective question directing students to think about theory change in terms of model change might have been helpful.

Although students did not show the same large gains in *theory change* as in the *changing nature of models*, what gains were achieved in *theory change* were much more strongly correlated to CTSR than were the gains in *changing nature of models*. In fact, the *changing nature of models* scores showed one of the weakest correlation between residualized gain and CTSR score of any category ($r = 0.15$) whereas *theory change* showed the strongest ($r = 0.41$). One possible explanation for this result might be that the those students with a higher CTSR score might have been more able to transfer this knowledge on model change to a similarly structured knowledge base (theory change), whereas those students with lower CTSR scores may have viewed theories and models as discrete entities and thus not seen the transfer of knowledge as necessary and applicable. Regardless of the reason, this correlation of CTSR score with theory change was significant when analyzed with a regression ($p = 0.001$).

In conclusion, it appears that the modeling approach was somewhat successful at teaching students about *theory change*, but this knowledge was transferred less well from the knowledge of the *changing nature of models* than was hoped. Furthermore, the normalized change was concentrated among those with the highest CTSR score. As knowledge of theory change was already significantly ($p = 0.034$) correlated with CTSR on the pretest, this becomes a case of the rich getting richer, and thus not the most appropriate way to teach students about theory change if the goal is No Child Left Behind.

Nature of hypotheses, theories and laws.

The data collected on the pretest and posttest indicates that this particular aspect of NOS was not improved over the course of the study. As a sub-score, questions one through six gave an pretest total average of 2.78 (out of 8.00), a posttest total of 2.95, resulting in an average normalized change of -0.01, a total Cohen's $d = 0.13$. Therefore, there was virtually no gain, and students exited class with the same poor understanding of the nature of hypotheses, theories and laws as they entered class with. Although there was some misunderstanding with question one and question six was long and had multiple parts (thus incomplete answers could be a result of forgetting to answer part of the question or because of ignorance), the data collected overwhelming points to no gain in student understanding of the nature of hypotheses, theories and laws. Students actually showed a negative normalized change and Cohen's d for questions one through three, a small positive (but not meaningful) change for questions four and five (with Cohen's d still negative), and a small positive Cohen's d and normalized change for question six. Even looking only at the free-response question (question 6), although the effect size was moderate (Cohen's $d = 0.52$), the gain was from a mean of 0.53 on the pretest to a mean of 0.97 on the posttest out of 3.00 possible points. This gain was not universal, but rather approximately 1/3 of the class improved

by one or two points, and the bulk of the class remained unchanged. These small gains were primarily centered around clarification of hypotheses and theories, not surprising given the class emphasis on forming hypotheses and the abovementioned link between models and theories.

Although there were virtually no gains, this category showed that the residualized gains were moderately correlated ($r = 0.31$) with CTSR score and this correlation was significant ($p = .015$). Specifically, what that means with this question is that often, those with low CTSR scores more frequently showed a negative normalized change, while those with a higher CTSR score more frequently showed a small positive normalized change, resulting in overall nearly no change but still a correlation. This result was seen with both the free-response and Likert-scale questions.

In conclusion, it has again been shown that in a category for which no explicit reflective questions are asked, small gains are observed. While the potential should be there to think about hypotheses, theories, and laws while modeling, these connections were not made. Obviously, like previous categories facing this issue, this category would benefit from additional study with explicit questions asking students to think about hypotheses, theories, and laws during modeling activities. One such way would be to have students look at models with great predictive power but with little explanation of how the numbers were calculated (one carbon footprint model asked about diet, wardrobe, and banking, which did not seem to be very related to carbon footprint at first). These models are in some way are more like laws, they say what will be observed, but not why. These law-like models could be contrasted with theory-like models, with better explanations of why certain variables were related. Better explanations does not mean that the model itself is better, the model asking about diet and banking could in fact have provided estimates of carbon footprint that were much more accurate than those of the theory-like model. It also does not mean either model could not be changed. A section like this, added to the carbon footprint model, might

have been capable of changing student's perceptions of *hypotheses, theories, and laws* more effectively than the current study.

Scientific method.

The final category related to the Nature of Science was the category *scientific method*. Likert-scale questions 40-43 (one point) and free response question 44 (3 points) comprised this category. Students appeared to have little difficulty understanding the questions. The average pretest score was 3.36 (out of 7.00) with a raw gain of 0.46 pushing the posttest to an average of 3.82. This raw gain translated into a small normalized change (0.12) and Cohen's $d = 0.39$.

In addition to showing only small gains, this category's residualized gains showed only a small correlation ($r = 0.2$) with CTSR. This correlation was not significant ($p = .124$).

While this section is worthy of more analysis, other than the idea that models are one means of scientific investigation, this question was not discussed explicitly in class and was tangentially related at best. Therefore, no further analysis was performed. As a whole, questions 40 to 44 represent a section of the test that could have been omitted. Although small gains were seen in some questions, overall, this section was not closely enough related to the classroom activities to merit inclusion.

Conclusion.

In conclusion, the results from each category individually point to a less successful conclusion for research sub-question three than with either sub-question one or sub-question two. Cohen's *d*'s were small and variable, with the best, theory change, being the area most closely associated with the actual activities in class, and the other two areas with small or no correlations. As none of these areas was as explicitly reflected upon as the modeling sections, this lack of explicit reflection (along with lack of repetition) remains the most obvious explanation for lack of

universal gain. On the other hand, correlations between residualized gain and CTSR scores were moderate in two of the three categories, and small in the third. This was interpreted as being related to the ability of high formal students being better able to transfer knowledge from models to theories.

The results of this sub-question are important because this study was initially started to help classroom teachers with a novel approach to teaching the nature of science, through models. Certainly, in its current form, a teacher would not be advised to use this approach as the approach shows little positive gain on student scores related to the Nature of Science, and what gain is shown, is concentrated among those that have the highest CTSR scores and are already doing well in NOS. On the other hand, it is hypothesized that with a more explicit reflective methodology, student gains in these NOS sections, as well as the two lower scoring modeling categories (*exact replicas* and *explanatory tools*) would also show stronger gains.

Overall Conclusion

The overall conclusion to the research question, as viewed through the sub-questions is that where the methodology was implemented correctly, with explicit reflective questions (such as most of the modeling categories), students showed large gains that were well in excess of the *a priori* Cohen's *d* of 0.5. On the other hand, the nature of science questions lacked this explicit reflection (as did the categories of models as *exact replicas* and models as *explanatory tools*, by and large) and showed much smaller gains. On the other hand, these categories were not explicitly reflected on because they were not the primary focus of the class activities. Therefore, where the methodology was executed correctly, the expected gains were shown, but in the other categories where gain was not shown, it is not known whether explicit reflection using existing activities or new activities more closely related to the stated learning goals (or a combination of both) would

have been needed to secure larger gains. Overall, the approach used for modeling instruction in this study appears to be an effective method for teaching modeling, but not an effective method for teaching NOS topics.

Likewise, the correlation between CTSR and the normalized changes on the various sub-scores on the pretest and posttest are widely varied. In four cases, the correlation is moderate, in four cases, small, and in two cases, there was no correlation. Furthermore, a ceiling effect appeared to be problem in at least one case. Similar results were observed across the classroom activities, although the actual modeling project itself tended to show consistently stronger relationships between CTSR score and success than other activities. Therefore, it is cautiously concluded that there is a link between cognitive development as measured by CTSR scores and at least certain aspects of the knowledge of models, and particularly between cognitive development and the ability to model.

Implications.

There is some evidence to suggest a certain level of cognitive development is preferable, if not quite necessary, to understand and build models. There is evidence that teaching a content course using models can increase modeling knowledge, and that this gain is also related to cognitive development. Therefore, it is recommended that modeling instruction be integrated into science content lessons. On the other hand, this approach to teaching modeling does not directly lead to gains in nature of science concepts, particularly for those students of low cognitive development, so nature of science concepts need to be reinforced explicitly.

When modeling activities are placed in curriculum, conventional wisdom would interpret the results of this study to indicate that care should be taken not to place activities in grades where the majority of students would be less likely to find success due to lower cognitive development.

Using this train of logic, modifying existing models and comparing multiple models of the same phenomenon both would make sense at lower grades and/or lower cognitive development. Likewise, creating models de novo and using models to make hypotheses might best be saved for later grades/higher cognitive development.

On the other hand, if this approach is followed too its conclusion, it would seem to lead to a situation where students are not taught in their zone of proximal development, but below it, stagnating further development. Some would argue that it has been this approach of not cognitively challenging secondary students that leads to great numbers of college age students who have not reached the full formal stage of cognitive development. Modeling activities would seem to lend themselves well to developing formal thought. By inserting modeling activities at an earlier grade, where students are developing formal thought instead of waiting for students to have developed formal thought, perhaps modeling activities could be used to accelerate this development. Adey and Shayer (1990) used cognitively challenging tasks to increase cognitive development in the students that they studied (see also Shayer and Adey (1992a, 1992b, 1993).

Opportunities for further research.

The results of this study yield more questions than answers. These questions offer several opportunities for further study, at least two of which stem from methodology flaws and another that should help clarify some of the relationships. Specifically, these opportunities are to improve the instrument, to improve the methodology and to repeat the study on a younger sample.

First, the pretest/posttest instrument should be revised again. It is now clear that in settling for an instrument available in the literature rather than going through the process of creation and validation of a new instrument, this research almost failed to provide interpretable results in some areas. Several of the questions from the SUMS should be revised further or eliminated.

Specifically, all questions containing the word *or* should be split into separate questions or eliminated. In addition, several other questions were identified in the follow-up interviews as having confusing wording, and this wording should be changed, even if the SUMS question from which it was built is no longer recognizable in the final version. Finally, question distribution across sub-scores needs to be more balanced. Additional questions on *how models are created, types of models,* and *changing nature of models* in particular should be added so that each has a balance of Likert-scale and free response to give more spread to the category sub-scores, which should help with both ceiling effects and regression analysis. It is also possible that some of the NOS questions should be removed unless the methodology is changed.

In addition to changing the pretest/posttest, it would be necessary to change the methodology to include explicit reflective questions for all categories for which this had not been done in this study, to rule out whether or not it was the lack of reflective questions or the lack of repetition that resulted in the low gains in these five categories. As the initial idea was to teach NOS through models, it is still preferable to add reflective NOS questions to the modeling activities, rather than add additional NOS activities that would be contrary to the idea of the research.

Finally, repeating this study with a younger sample would offer several advantages. First, this sample would in some ways be more homogenous, having more similar backgrounds in the number of science and math classes taken, age, etc. Eliminating these variables would perhaps help to focus more control on studying the effect of cognitive development on gain. In addition, a younger sample should provide a greater number of students below the CTSR score threshold of 14.5 which defines the lower end of formal operational developmental level, and for questions that suffered from too high a success rate, a younger sample might achieve more balance. If there is a

relationship between CTSR score and gain, then comparing the results of this younger group with the results from the older group would provide an interesting contrast.

REFERENCES

Abd-El-Khalick, F., Bell, R., & Lederman, N. G. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education, 82*(4), 417-436.

Abd-El-Khalick, F., Lederman, N. G., Bell, R. L., & Schwartz, R. S. (2001, January). Views on nature of science questionnaire (VNOS): Toward valid and meaningful assessment of learners' conceptions of nature of science. *Proceedings of the Annual Meeting of the Association for the Education of Teachers in Science*, Costa Mesa, California.

Adey, P.S., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school pupils. *Journal of Research in Science Teaching*, *27*(3), 267-285.

Aikenhead, G., & Ryan, A. (1992). The development of a new instrument: "Views on Science-Technology-Society" (VOSTS). *Science Education, 76*(5), 477-491.

Akerson, V. L., Abd-El-Khalick, F., & Lederman, N. G. (2000). Influence of a reflective explicit activity-based approach on elementary teachers' conceptions of nature of science. *Journal of Research in Science Teaching, 37*(4), 295–317.

American Association for the Advancement of Science. (1989). *Science for all Americans.* Oxford: Oxford University Press.

Asami, N., King, J., & Monk, M. (2000). Tuition and memory: Mental models and cognitive processing in Japanese children's work on d.c. electrical circuits. *Research in Science & Technology Education, 18*(2), 141-154.

Bliss, J. (1994). From mental models to modelling. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, and C. Tompsett (Eds), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Boohan, R. (1994). Interpreting the world with numbers: An introduction to quantitative modeling. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, & C. Tompsett (Eds.), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Campbell, R., & Olson, D. R. (1990). Children's thinking. In R. Grieve & M. Hughes (Eds.), *Understanding children: Essays in honour of Margaret Donaldson* (pp. 189-209). Oxford: Blackwell.

Carey, S., Evans, R., Honda, E.J., & Unger, C. (1989). An experiment is when you try it and see if it works: A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*, 514-529.

Cartier , J, Rudolph, J, & Stewart, J, (2001). The Nature and Structure of Scientific Models. *The National Center for Improving Student Learning and Achievement*, Retrieved October 18, 2006, from http://www.wcer.wisc.edu/ncisla/publications/reports/Models.pdf.

Center for Disease Control. (2007). *National health and nutrition examination survey – United States growth charts - Data files*. Retrieved January 11, 2007, from http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/datafiles.htm.

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education, 22*(9), 1041–1053.

Clement, J., & Steinberg, M. (2002). A step-wise evolution of mental models of electric circuits: A "learning-aloud" case study. *The Journal of the Learning Sciences,11*(4), 389–452.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chittleborough, G., Treagust, D., Mocerino, M., & Thapelo, M. (2005). Students' perceptions of the role of models in the process of science and in the process of learning, *Research in Science and Technological Education, 23*(2),195-212

Committee on Conceptual Framework for the New K-12 Science Education Standards, National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

Cresswell, J. W. (2003). *Research Design.* Thousand Oaks, CA: Sage Publications.

Cullin, M., & Crawford, B. (2003). Using technology to support prospective science teachers in learning and teaching about scientific models. *Contemporary Issues in Technology and Teacher Education* [online serial], *2*(4). Available: http://www.citejournal.org/vol2/iss4/science/article1.cfm.

Cullin, M. (2004). Examining perspective science teachers' understanding of the role of models and modeling in science within the context of building and testing computer models of pond ecosystems. Doctoral dissertation. The Pennsylvania State University.

Dienes, Z. (1960). *Building Up Mathematics* (4th edition). London: Hutchinson Educational Ltd.

Driscoll, M. P. (1994). *Psychology of learning for instruction.* Needham Heights, MA: Allyn & Bacon.

Driver, R. (1978). When is a stage not a stage: A critique of Piaget's theory of cognitive development and its application to science education. *Educational Research, 21*(1), 54-61.

Ergazaki, M., Komis, V., & Zogza, V. (2005). High-school students' reasoning while constructing plant growth models in a computer-supported educational environment. *International Journal of Science Education 27*(8) 909–933.

Forbus, K., Carney, K., Sherin, B., & Ureel, L. (2004, August). Qualitative modeling for middle-school students. *Proceedings of the 18th International Qualitative Reasoning Workshop,* Evanston, IL, USA. Available: http://www.qrg.northwestern.edu/people/ureel/papers/QR04_VModel_Final.pdf.

Gill, I. (2007, September 14). Minnesota State University Moorhead Fact Book. Retrieved February 29, 2008, from Minnesota State University Web site: http://www.mnstate.edu/institut/fact_book_home_page/msum_fact_book.htm.

Gove, P. (1981). Webster's Third New International Dictionary of the English Language, Unabridged. Springfield, MA: Merriam-Webster, Inc.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28,* 799–822.

Gutwill, J., Frederiksen, J., & White, B. (1999). Making their own connections: Students' understanding of multiple models in basic electricity. *Cognition & Instruction*, *17*(3), 249 – 282.

Harrison, A. (1998). Modeling science lessons: Are there better ways to learn with models? *School Science & Mathematics*, *98*(8), 420-429.

Harrison, A. & Treagust, D. (2000). Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry. *Science Education*, *84*(3), 352-381.

Harrison, A.,G. (2001, March). *Models and PCK: Their relevance for practicing and pre-service teachers.* Paper presented at the annual meeting of the National Association of Research in Science Teaching, St. Louis, MO.

Hooft, G. (2007). The making of the standard model. *Nature, 448(*19), 271-273.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference and consciousness.* Cambridge University Press.

Justi, R., & Gilbert, J. (2002a). Modeling, teachers' views on the nature of modeling, and implications for the education of modelers. *International Journal of Science Education, 24*(4), 369-387.

Justi, R., & Gilbert, J. (2002b). Science teachers' knowledge about and attitudes towards the use of models and modeling in learning science. *International Journal of Science Education, 24*(12), 1273-1292.

Kehle, P., & Lester, F. (2003). A semiotic look at modeling behavior. In R. Lesh & H. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning and teaching*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Krathwohl, D. R. (1998). *Methods of Educational and Social Science Research: An Integrated Approach*. Long Grove, IL: Waveland Press.

Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11–24 .

Lawson, A. E., Alkoury, S., Benford, R., Clark, B., & Falconer, K. (2000). What kind of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching, 37*(9), 996-1018.

Lawson, A.E., Clark, B., Cramer-Meldrum, E., Falconer, K., Sequist, J., & Kwon, Y. (2000). Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist? *Journal of Research in Science Teaching, 37*(1), 81–101.

Lawson, A.E., Drake, N., Johnson, J., Kwon, Y., & Scarpone, C. (2000). How good are students at testing alternative hypotheses involving unseen entities? *The American Biology Teacher*, *62*(4), 249-255.

Lawson, A.E., Banks, D., & Logvin, M. (2007). Self-efficacy, reasoning ability, and achievement in college biology. *Journal of Research in Science Teaching, 44*(5), 706-24.

Lederman, N.G., Wade, P.F., & Bell, R.L. (1998). Assessing understanding of the nature of science: A historical perspective. In W. McComas (Ed.), *The nature of science in science education: Rationales and strategies* (pp. 331-350). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lehrer, R., & Romberg, T. (1996). Exploring Children's Data Modeling. *Cognition and Instruction 14*(1), 69-108.

Lehrer, R., & Schauble, L. (2003). Origins and evolution of model-based reasoning in mathematics and science. In R. Lesh & H Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning and teaching*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lesh, R. & Carmona, G. (2003). Piagetian conceptual systems and models for mathematizing everyday experiences. In R. Lesh & H Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning and teaching*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lesh, R., Cramer, K., Doerr, H., Post, T., & Zawojewski, J. (2003). Model development sequences. In R. Lesh & H Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning and teaching*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lesh, R., & Doerr, H. (2003). Foundations of models and modeling: Perspective on mathematics teaching, learning and problem solving. In R. Lesh & H Doerr (Eds.), *Beyond constructivism: models and modeling perspectives on mathematics problem solving, learning and teaching*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Liang, L., Chen, S., Chen, X., Kaya, O., Adams, A., Macklin, M., & Ebenezer, J. (2006, April). *Student Understanding of Science and Scientific Inquiry (SUSSI): Revision and Further Validation of an Assessment Instrument*. Paper presented at the annual conference of the National Association for Research in Science Teaching (NARST), San Francisco, CA.

Mellar, H., & Bliss, J. (1994). Modelling and education. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, and C. Tompsett (Eds), *Learning with artificial worlds: computer based modeling in the curriculum*. London: The Falmer Press,

Mishler, E. G. (1990). Validation in enquiry-guided research: The role of exemplars in narrative studies. Harvard Educational Review, 60, 415-442.

Minnesota Department of Education. (2005). MCA-II. Retrieved October 17, 2006, from Minnesota Department of Education Web site: http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MCA_II/index.html.

Minnesota Department of Education. (2006, August 10). MCA-II: Test Specifications for Science. Retrieved October 17, 2006, from Minnesota Department of Education Web site: http://education.state.mn.us/mde/static/006366.pdf.

Minnesota State University Moorhead. (2006, March 24). Dragon core: Competency areas. Retrieved August 16, 2007, from Minnesota State University Moorhead Web site: http://www.mnstate.edu/acadaff/dragoncore/CAs.htm.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. U.S. Department of Education, Washington, D.C.

National Committee on Science Education Standards and Assessment (NCSESA). (1996). *National Science Education Standards*. National Research Council, Washington D.C.

National Science Teachers Association. (2003, July*). NSTA position statement: The teaching of evolution*. Retrieved Feb 11, 2008, from National Science Teacher Association Web site: http://www.nsta.org/about/positions/evolution.aspx.

Next Generation Science Standards. (2012). Achieve, Inc. Retrieved November 15, 2012 from http://www.nextgenscience.org/three-dimensions.

Ogborn, J. (1994). Overview: The nature of modelling. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, and C. Tompsett (Eds), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Ogborn, J., & Mellar, H. (1994). Models: Their maker, uses, and problems. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, and C. Tompsett (Eds), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Ogborn, J., & Miller, R. (1994). Computational issues in modeling. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, and C. Tompsett (Eds.), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Penner, D., Lehrer, R., & Schauble, L. (1998). From physical models to biomechanics: A design-based modeling approach. *The Journal of the Learning Sciences*, *7*(3&4), 429-449.

Piaget, J., & Inhelder, B. (1955). The growth of logical thinking from childhood to adolescence. In H. E. Gruber & J. J. Voneche (Eds.), *The Essential Piaget* (p. 405-444). New York: Basic Books.

Piaget, J., & Inhelder, B. (1966). The preadolescent and the propositional operation. In H. E. Gruber & J. J. Voneche (Eds.), *The Essential Piaget* (p.394-444). New York: Basic Books.

Reif, F., & Larkin, J. (1991). Cognition in scientific and everyday domains: Comparison and learning implications. *Journal of Research in Science Teaching, 28*(9), 733-760.

Sakonidis, H. (1994). Representations and representation systems. In H. Mellar, J. Bliss, J. Bohan, J. Ogborn, & C. Tompsett (Eds), *Learning with artificial worlds: Computer based modeling in the curriculum*. London: The Falmer Press.

Sarri, H., & Viiri, J. (2003). A research-based teaching sequence for teaching the concept of modelling to seventh-grade students. *International Journal of Science Education,* 2*5*(11), 1333-1352.

Schoenfeld, A. H. (1982). Measures of problem solving performance and of problem solving instruction. *Journal of Research in Mathematics Education, 13*(1), 31-49.

Schwartz, R., & Lederman, N. (2005, April). *What scientists say: Scientists' views of models*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing student's understanding of scientific modeling. *Cognition and Instruction, 2*3(2), 165-205.

Shayer, M., & Adey, P.S. (1981). Towards a Science of Science Teaching. London: Heinemann Educational Books.

Shayer, M., & Adey, P.S. (1992a). Accelerating the development of formal thought in middle and high school students. III: Testing the permanency of effects. *Journal of Research in Science Teaching*, *29*(10), 1101–1115.

Shayer, M., & Adey, P.S. (1992b). Accelerating the development of formal thinking in middle and high school students. II: Post-project effects on science achievement. *Journal of Research in Science Teaching*, *29*(1), 81–92.

Shayer, M., & Adey, P.S. (1993). Accelerating the development of formal thinking in middle and high school students. IV: Three years on after a two year intervention. *Journal of Research in Science Teaching*, *30*(4), 351–366.

Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? *Journal of Research in Science Teaching, 28*(9), 761-784.

Sutherland, R., & Rojano, T. (1993). A spreadsheet approach to solving algebra problems. *Journal of Mathematical Behavior 12,*353-383.

Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Student's understanding of the role of scientific models in learning science. *International Journal of Science Education 24*(4), 357-368.

Trowbridge, L. W., Bybee, R. W., & Powell, J.C. (2000). *Teaching Secondary School Science* (7th ed.). Upper Saddle River, NJ: Merrill Publishing Company.

Turner, S. (2008). School science and its controversies; or, whatever happened to science literacy? *Public Understanding of Science*, *17*(1), 55-72.

Valanides, N., & Angeli, C. (2006). Preparing preservice elementary teachers to teach science through computer models. *Contemporary Issues in Technology and Teacher Education* [Online serial], *6*(1). Available: http://www.citejournal.org/vol6/iss1/science/article1.cfm

Van Driel, J., & Verloop. N. (1999). Teachers' knowledge of models and modeling in science. *International Journal of Science Education, 21*(11), 1141-1153.

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10*(1), 1-100.

Windschitl, M., & Thompson, J. (2006). Transcending simple forms of school science investigation: The impact of pre-service instruction on teachers' understandings of model-based inquiry. *American Educational Research Journal, 43*(4), 783-835.

Wisnudel-Spitulnik, M., Kracjik, J., & Soloway, E. (1999). Construction of models to promote scientific understanding. In W. Fuerzeig & N. Roberts, (Eds.), *Modeling and Simulation in Science and Mathematics Education (p. 70-94*). New York: Springer-Verlag.

APPENDIX A: INSTRUMENTS.

Nature of Science and Modeling Pretest and Posttest.
Final draft. Adapted from SUMS and SUSSI tests.

1. Scientific theories exist in the natural world and are uncovered through scientific investigations.	S.D. Theories are created by scientists to explain the natural world.
2. Unlike theories, scientific laws are not subject to change.	S.D. Scientific laws are subject to change Newton's Laws of motion do not hold at relativistic speeds (although students have less experience with laws changing than theories, as most examples are found in modern physics).
3. Scientific laws are theories that have been proven.	S.D. Theories and laws answer different questions. Laws tell what phenomenon will be observed, often with great accuracy, but theories postulate why.
4. Scientific theories explain scientific laws.	S.A.
5. Scientific theories are hypotheses that have been tested many times and not disproven.	S.A. Some hypotheses become theories through repeated testing.

6. With examples where appropriate, what is the nature (definition) of each: law, hypothesis, and theory. Then, explicitly state the differences and relationships between each.	Hypothesis – a testable prediction Theory – the best current explanation of a related phenomena Law – A well-tested, typically mathematical, relationship between a number of variables. Hypotheses that are supported can become parts of theories or laws. Theories do NOT become laws, contrary to student beliefs, but may explain them.
7. Scientific theories are subject to on-going testing and revision.	S.A.
8. Scientific theories may be completely replaced by new theories in light of new evidence.	S.A.
9. Scientific theories may be changed because scientists reinterpret existing observations.	S.A.
10. Scientific theories based on accurate experimentation will not be changed.	S.D. A theory may correctly explain all "accurate" experimentation that exist at that time, and yet still be changed as new data become available.

11. Do scientific theories change? If yes – how (in what ways and to what extent) and why? If no – why not?	Yes. The theories may change gradually or radically based on new evidence.
12. List as many scientific models as you can.	A variety of models should be represented including physical, mathematical, and conceptual/theoretical models.
13. Multiple models exist of the same phenomenon, such as a map of the United States. Why?	Different models reflect different aspects (roads, political boundaries, geography) of the same phenomenon (the United States). Each serves a different purpose.
14. What is the most important characteristic of a scientific model or, in other words, what characteristic makes a scientific model the most useful? Explain.	The ability to make accurate, testable hypotheses. To adequately explain a variety of observations.
15. A headline reads "Global warming model predicts sea-level will rise 2 meters by 2100 A.D.". What do they mean by "model" and how was this model created?	This mathematical model was likely physically created on a computer by conscious choice of the variables and data to include and omit.
16. Scientific models are only used to physically or visually represent something.	S.D. Mathematical, conceptual, or theoretical models may not be physical or visual.
17. Scientific models are used to explain scientific phenomena.	S.A.

18. Scientific models may be used to show an idea.	S.A.
19. A scientific model is a diagram, picture, map, graph or photo of a physical object.	S.D. Most models used in science for investigations are not of physical objects, but rather of relationships.
20. Models are used to help formulate ideas and theories about scientific events.	S.A.
21. Scientific models' primary value is in showing/teaching science.	S.D. Models' primary value lies in their ability to make accurate predictions.
22. Models are used to make and test predictions about a scientific event.	S.A.
23. A model can change if new theories or evidence prove otherwise.	S.A.
24. Once created, a model does not change.	S.D. Models change with new ideas, theories, or experimental data.
25. A model can change if there are changes in data or beliefs.	S.A.
26. Multiple models of the same phenomenon/object are typically used to express features of a phenomenon/object by showing different perspectives to view/see a phenomenon/object.	S.D. Multiple models of the same phenomenon tend to show different interactions a phenomenon may make, rather than different views/perspectives of how an object looks.
27. Multiple models of the same	S.A.

phenomenon/object represent different versions/aspects/facets of the phenomenon/object.	
28. Models can show the relationship of ideas clearly.	S.A.
29. Multiple models of the same phenomenon/object are used to show differences in individual's theories on what things look like and/or how they work.	S.A.
30. Multiple scientific models are used primarily to show different sides or shapes of an object.	S.D. Many scientific models are not physical.
31. Multiple models of the same object/phenomenon may use different information.	S.A.
32. A model has what is needed to show or explain a scientific phenomenon.	S.A.
33. A scientific model should be an exact replica of the object.	S.D. If a model were an exact replica, it would no longer be a model, it would be the original.
34. A model needs to accurately represent the object/phenomenon in the areas of interest.	S.A.

35. A model should closely resemble the object/phenomenon, so nobody can disprove it.	S.D. The utility and thus longevity of a model depends more on its ability to functionally represent the phenomenon, not the apparent physical similarity.
36. All parts of a model should have an understandable purpose/reason.	S.A.
37. A scientific model needs to be close to the real thing by being very exact in every way except for size.	S.D. Since many scientific models are NOT physical, most are not scale models.
38. A model shows what the real thing does and/or what it looks like.	S.A.
39. Multiple models are important for different student learning styles.	S.D. Contrary to some students' beliefs, multiple models do NOT have anything to do with learning styles.
40. Scientists use different types of methods to conduct scientific investigations.	S.A.
41. Scientists follow the same step-by-step scientific method.	S.D. See below.
42. Correct use of the scientific method guarantees accurate results.	S.D. The scientific method does not automatically eliminate random or systematic error.
43. Experiments are not the only means	S.A. See below.

used in the development of scientific knowledge.	
44. With examples, explain whether scientists follow a single, universal scientific method OR use different types of methods.	Astronomy or field biology may be purely observational of natural phenomena. Other sciences may use strictly controlled experiments. Different methods are valid for different disciplines.

Screen Capture of the Pretest/Posttest.

What the students will see when they take this instrument. The text boxes are unlimited.

Please Note: It is recommended that you save your response as you complete each question.

Question 1 (1 point) Save

Scientific theories exist in the natural world and are uncovered through scientific investigations.

SA A N D SD

Question 2 (1 point) Save

Unlike theories, scientific laws are not subject to change.

SA A N D SD

Question 3 (1 point) Save

Scientific laws are theories that have been proven.

SA A N D SD

Question 4 (1 point) Save

Scientific theories explain scientific laws.

SA A N D SD

Question 5 (1 point) Save

Scientific theories are hypotheses that have been tested many times and not disproven.

SA A N D SD

Question 6 (3 points) Save

With examples where appropriate, what is the nature (definition) of each: law, hypothesis, and theory. Then, explicitly state the differences and relationships between each.

Question 7 (1 point) Save

Scientific theories are subject to on-going testing and revision.

SA A N D SD

Question 8 (1 point) Save

Scientific theories may be completely replaced by new theories in light of new evidence.

SA A N D SD

Question 9 (1 point) Save

Scientific theories may be changed because scientists reinterpret existing observations.

SA A N D SD

Question 10 (1 point) Save

Scientific theories based on accurate experimentation will not be changed.

SA A N D SD

Question 11 (3 points) Save

Do scientific theories change? If yes – how (in what ways and/or to what extent) and why? If no – why not?

Question 12 (3 points) Save

List as many scientific models as you can.

Question 13 (3 points) Save

Multiple models exist of the same phenomenon, such as a map of the United States. Why?

Question 14 (3 points) Save

What is the most important characteristic of a scientific model or, in other words, what characteristic makes a scientific model the most useful? Explain.

Question 15 (3 points) Save

A headline reads "Global warming model predicts that sea level will rise 2 meters by 2100 AD". What do they mean by "model" and how was this model created?

Question 16 (1 point) Save

Scientific models are only used to physically or visually represent something.

Strongly Disagree Disagree Not Sure Agree Strongly Agree

Question 17 (1 point) Save

Scientific models are used to explain scientific phenomena.

Strongly Disagree Disagree Not Sure Agree Strongly Agree

Question 18 (1 point) Save

Scientific models are used to show an idea.

Strongly Disagree Disagree Not Sure Agree Strongly Agree

Question 19 (1 point) Save

A scientific model is a diagram, picture, map, graph or photo of a physical object.

Strongly Disagree Disagree Not Sure Agree Strongly Agree

Question 20 (1 point) Save

Models are used to help formulate ideas and theories about scientific events.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 21 (1 point) Save

Scientific models primary value is in showing/teaching science.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 22 (1 point) Save

Models are used to make and test predictions about a scientific event.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 23 (1 point) Save

A model can change if new theories or evidence prove otherwise.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 24 (1 point) Save

Once created, a model does not change.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 25 (1 point) Save

A model can change if there are changes in data or beliefs.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 26 (1 point) Save

Multiple models of the same phenomenon/object are typically used to express features of a phenomenon/object by showing different perspectives to view/see a phenomenon/object.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 27 (1 point) Save

Multiple models of the same phenomenon/object represent different versions/aspects/facets of the phenomenon/object.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 28 (1 point) Save

Models can show the relationship of ideas clearly.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 29 (1 point) Save

Multiple models of the same phenomenon/object are used to show differences in individual's theories on what things look like and/or how they work.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 30 (1 point) Save

Multiple scientific models are used primarily to show different sides or shapes of an object.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 31 (1 point) Save

Multiple models of the same object/phenomenon may use different information.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 32 (1 point) Save

A model has what is needed to show or explain a scientific phenomenon.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 33 (1 point) Save

A scientific model should be an exact replica of the object.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 34 (1 point) Save

A model needs to accurately represent the object/phenomenon in the areas of interest.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 35 (1 point) Save

A model should closely resemble the object/phenomenon, so nobody can disprove it.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 36 (1 point) Save

All parts of a model should have an understandable purpose/reason.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 37 (1 point) Save

A scientific model needs to be close to the real thing by being very exact in every way except for size.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 38 (1 point) Save

A model shows what the real thing does and/or what it looks like.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 39 (1 point) Save

Multiple scientific models are important for different student learning styles.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Question 40 (1 point) Save

Scientists use different types of methods to conduct scientific investigations.

SA | A | N | D | SD

Question 41 (1 point) Save

Scientists follow the same step-by-step scientific method.

SA | A | N | D | SD

Question 42 (1 point) Save

Correct use of the scientific method guarantees accurate results.

SA | A | N | D | SD

Question 43 (1 point) Save

Experiments are not the only means used in the development of scientific knowledge.

SA | A | N | D | SD

Question 44 (3 points) Save

With examples, explain whether scientists follow a single, universal scientific method OR use different types of methods.

Save All Responses Go to Submit Quiz

Unscored copies of 20% of the pretests will be mixed in with the post test, free response answers. This is to prevent an unconscious bias towards inflating the posttest scores, in order to show more gain. If the posttest scorings of the pretest are higher than their first scoring, a scoring bias is present.

Classroom Test of Scientific Reasoning

On the pages that follow is the Classroom Test of Scientific Reasoning used. It has not been modified in any way.

CLASSROOM TEST OF

SCIENTIFIC REASONING

Multiple Choice Version

Directions to Students:

This is a test of your ability to apply aspects of scientific and mathematical reasoning to analyze a situation to make a prediction or solve a problem. Make a dark mark on the answer sheet for the best answer for each item. If you do not fully understand what is being asked in an item, please ask the test administrator for clarification.

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

Revised Edition: August 2000 by Anton E. Lawson, Arizona State University. Based on: Lawson, A.E. 1978. Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1): 11-24.

1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*

   a. The pancake-shaped piece weighs more than the ball
   b. The two pieces still weigh the same
   c. The ball weighs more than the pancake-shaped piece

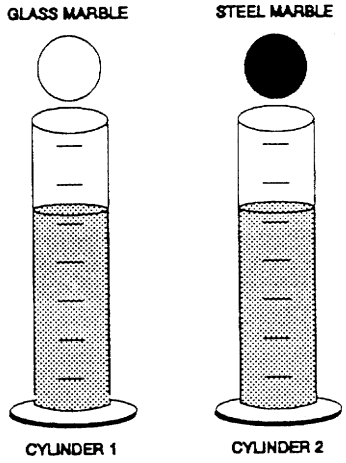
2. *because*

   a. the flattened piece covers a larger area.
   b. the ball pushes down more on one spot.
   c. when something is flattened it loses weight.
   d. clay has not been added or taken away.
   e. when something is flattened it gains weight.

3. To the right are drawings of two cylinders filled to the same level with water. The cylinders are identical in size and shape.

   Also shown at the right are two marbles, one glass and one steel. The marbles are the same size but the steel one is much heavier than the glass one.

   When the glass marble is put into Cylinder 1 it sinks to the bottom and the water level rises to the 6th mark. *If we put the steel marble into Cylinder 2, the water will rise*

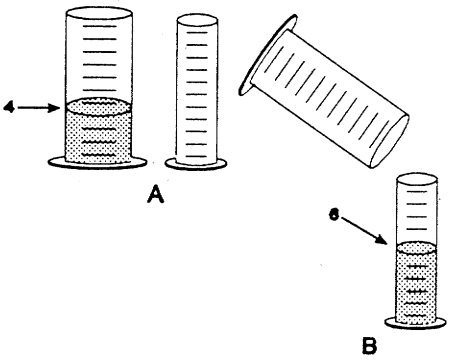


   a. to the same level as it did in Cylinder 1
   b. to a higher level than it did in Cylinder 1
   c. to a lower level than it did in Cylinder 1

4. *because*

   a. the steel marble will sink faster.
   b. the marbles are made of different materials.
   c. the steel marble is heavier than the glass marble.
   d. the glass marble creates less pressure.
   e. the marbles are the same size.

5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

a. to 8
b. to 9
c. to 10
d. to 12
e. none of these answers is correct

6. *because*

a. the answer cannot be determined with the information given.
b. it went up 2 more before, so it will go up 2 more again.
c. it goes up 3 in the narrow for every 2 in the wide.
d. the second cylinder is narrower.
e. one must actually pour the water and observe to find out.

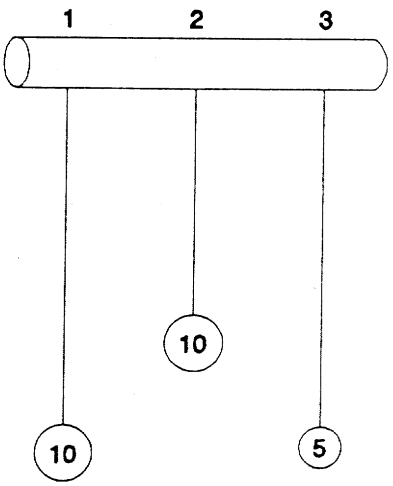
7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

a. to 9
b. to 8
c. to 71/2
d. to 7 1/3
e. none of these answers is correct

8. *because*

a. the ratios must stay the same.
b. one must actually pour the water and observe to find out.
c. the answer cannot be determined with the information given.
d. it was 2 less before so it will be 2 less again.
e. you subtract 2 from the wide for every 3 from the narrow.

9. At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10-unit weight is attached to the end of String 1. A 10-unit weight is also attached to the end of String 2. A 5-unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.



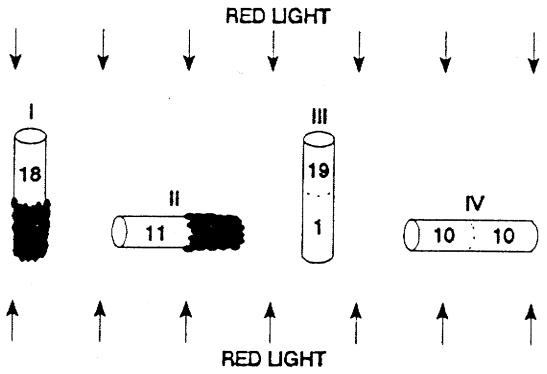
Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. *Which strings would you use to find out?*

a. only one string
b. all three strings
c. 2 and 3
d. 1 and 3
e. 1 and 2

10. *because*

a. you must use the longest strings.
b. you must compare strings with both light and heavy weights.
c. only the lengths differ.
d. to make all possible comparisons.
e. the weights differ.

11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



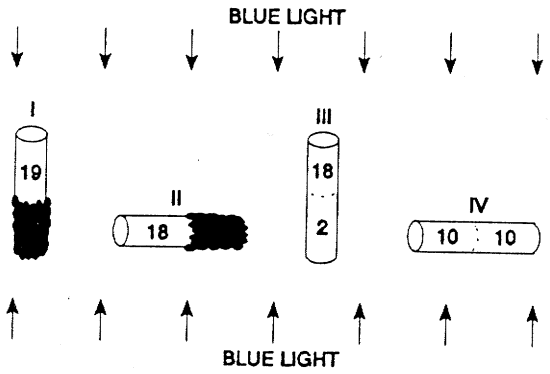
This experiment shows that flies respond to (respond means move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

12. *because*

a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
b. most flies did not go to the bottom of Tubes I and III.
c. the flies need light to see and must fly against gravity.
d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
e. some flies are in both ends of each tube.

13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.



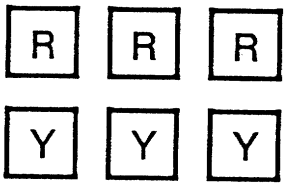
These data show that these flies respond to (respond means move to or away from):

a. blue light but not gravity
b. gravity but not blue light
c. both blue light and gravity
d. neither blue light nor gravity

14. *because*

a. some flies are in both ends of each tube.
b. the flies need light to see and must fly against gravity.
c. the flies are spread about evenly in Tube IV and in the upper end of Tube III.
d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
e. most flies are in the upper end of Tube I and the lighted end of Tube II.

15. Six square pieces of wood are put into a cloth bag and mixed about. The six pieces are identical in size and shape, however, three pieces are red and three are yellow. Suppose someone reaches into the bag (without looking) and pulls out one piece. *What are the chances that the piece is red?*

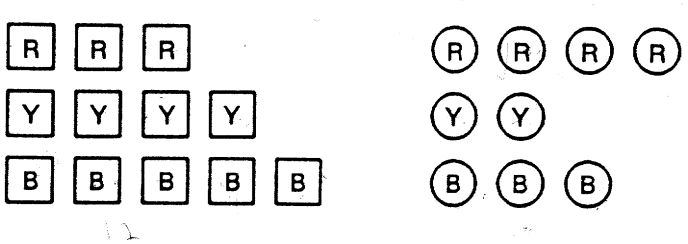


a. 1 chance out of 6
b. 1 chance out of 3
c. 1 chance out of 2
d. 1 chance out of 1
e. cannot be determined

16. *because*

a. 3 out of 6 pieces are red.
b. there is no way to tell which piece will be picked.
c. only 1 piece of the 6 in the bag is picked.
d. all 6 pieces are identical in size and shape.
e. only 1 red piece can be picked out of the 3 red pieces.

17. Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece.



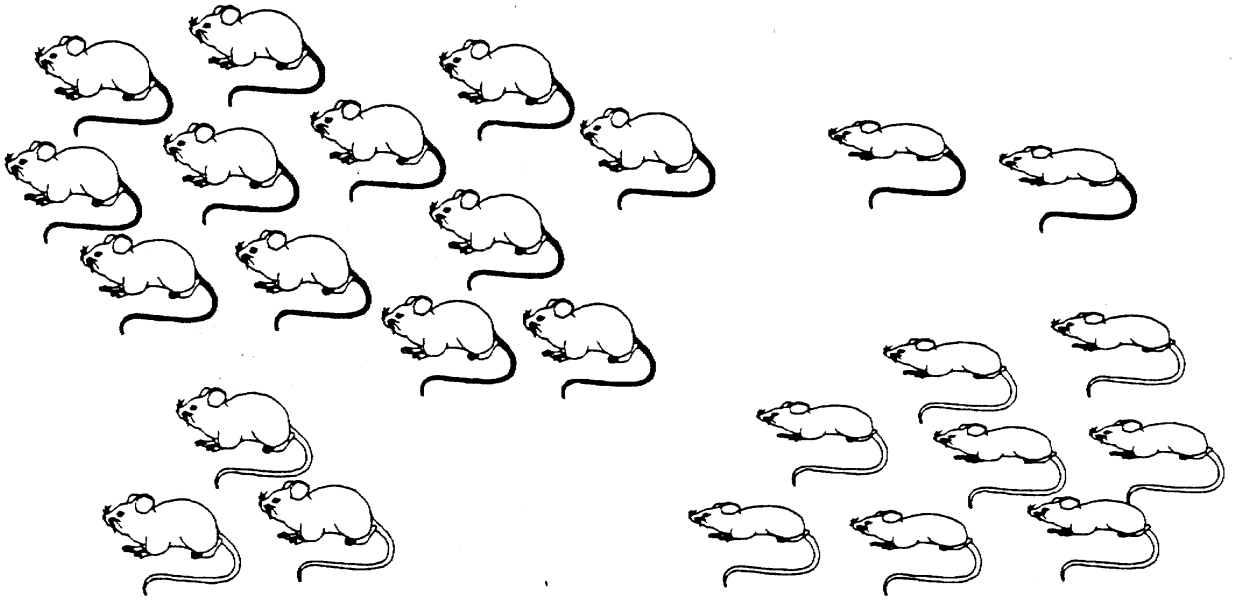
What are the chances that the piece is a red round or blue round piece?

a. cannot be determined
b. 1 chance out of 3
c. 1 chance out of 21
d. 15 chances out of 21
e. 1 chance out of 2

18. *because*

a. 1 of the 2 shapes is round.
b. 15 of the 21 pieces are red or blue.
c. there is no way to tell which piece will be picked.
d. only 1 of the 21 pieces is picked out of the bag.
e. 1 of every 3 pieces is a red or blue round piece.

19. Farmer Brown was observing the mice that live in his field. He discovered that all of the mice were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. Below are the mice that he captured.



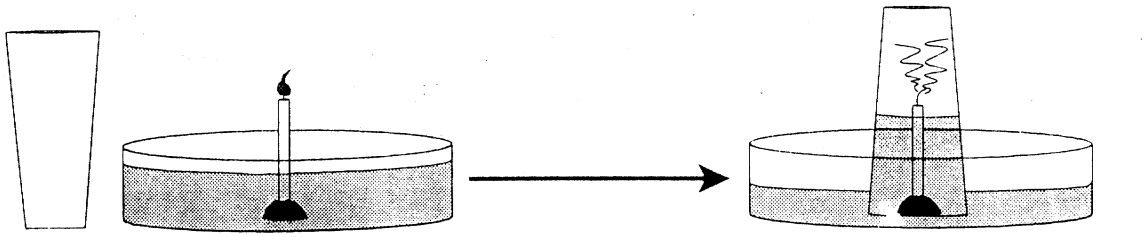
Do you think there is a link between the size of the mice and the color of their tails?

a. appears to be a link
b. appears not to be a link
c. cannot make a reasonable guess

20. *because*

a. there are some of each kind of mouse.
b. there may be a genetic link between mouse size and tail color.
c. there were not enough mice captured.
d. most of the fat mice have black tails while most of the thin mice have white tails.
e. as the mice grew fatter, their tails became darker.

21. The figure below at the left shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle, and placed in the water, the candle quickly goes out and water rushes up into the glass (as shown at the right).



This observation raises an interesting question: Why does the water rush up into the glass?

Here is a possible explanation. The flame converts oxygen into carbon dioxide. Because oxygen does not dissolve rapidly into water but carbon dioxide does, the newly formed carbon dioxide dissolves rapidly into the water, lowering the air pressure inside the glass.

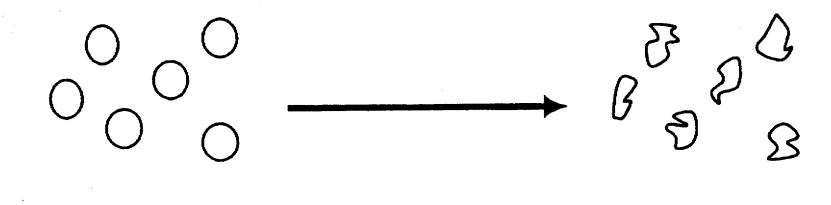
Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). *Using some or all of the materials, how could you test this possible explanation?*

a. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
b. The water rises because oxygen is consumed, so redo the experiment in exactly the same way to show water rise due to oxygen loss.
c. Conduct a controlled experiment varying only the number of candles to see if that makes a difference.
d. Suction is responsible for the water rise, so put a balloon over the top of an open-ended cylinder and place the cylinder over the burning candle.
e. Redo the experiment, but make sure it is controlled by holding all independent variables constant; then measure the amount of water rise.

22. What result of your test (mentioned in #21 above) would show that your explanation is probably wrong?

a. The water rises to the same level as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.

23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.



Magnified Red Blood Cells After Adding Salt Water

This observation raises an interesting question: Why do the red blood cells appear smaller?

Here are two possible explanations: I. Salt ions (Na+ and Cl-) push on the cell membranes and make the cells appear smaller. II. Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag, placing it in a salt solution for ten minutes, and then reweighing the bag.

What result of the experiment would best show that explanation I is probably wrong?

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

24. *What result of the experiment would best show that explanation II is probably wrong?*

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

APPENDIX B:

ACHIEVING INTER-RATER RELIABILITY

The rubrics used to score the SUSSIE pretest and posttest were developed by the author using data from the pilot study. Overall, there appeared to be some history with this group of students, as a number of students continuously mentioned the Middle East and/or India in their free response answers. One of these students mentioned the fact that belief systems in different countries were different came from another class she was taking at that time.

The two raters first developed a rubric without looking at the responses and rated the first three SUSSI free response questions. There was very low inter-rater reliability initially ($\rho = 0.1207, 0.3624, 0.1369$). After discussing these answers and amending the rubrics, consensus was reached on these scores, and changes regarding the wording of the question in future administrations of this exam were made. The rubrics for the next three SUSSI free response questions were created after looking over student answers and the responses were scored independently. The inter-rater reliabilities in this case were much higher ($\rho = 0.7489, 0.6215$, and 0.5058), all of which were significant at the $p = .01$ level. Finally, the raters created rubrics for the four free response modeling questions (59-62), scored these independently, and achieved $\rho = 0.3866, 0.5766, 0.3711, 0.5744$.

Question Five.

"With examples, explain why you think scientists' observations and interpretations are the same OR different."

This question posed a number of problems. First, the initial inter-rater reliability was very low (ρ [24] = .12, p = .57). From a student answer standpoint, virtually no students included examples in their answers. Since a lack of answer is ambiguous (did the student forget to answer the question or was the student unable to answer the question and thus left it blank) the question was modified to make the example portion explicit.

What the question is supposed to determine (from SUSSI taxonomy): "Science is based on both observations and inferences. Observations are descriptive statements about natural phenomena that are directly accessible to human senses (or extensions of those senses) and about which observers can reach consensus with relative ease. Inferences are interpretations of those observations. Perspectives of current science and the scientist guide both observations and inferences. Multiple perspectives contribute to valid multiple interpretations of observations."

Essential changes or clarifications:

The most applicable part of the above taxonomy to this study is how perspectives (such as a model or theory) guide observations and inferences. A number of student answers (three of 24) emphasized the difference between inference and observation. Four student answers hinted that two scientists could have different interpretations and observations if they examined different experiments. The final rubric is shown in Table 16.

Table 16. Final rubric for question 5.

Points	*Description*
3	Both a) Point of view/previous background knowledge/model/theory could influence observations and/or influences, therefore b) Different points of view could yield different observations and or inferences.
2	1 of the 2 points mentioned above
1	Inferences can be different (no explanation of how or why)
0	Observations are facts or equivalent statement.

To further delimit the range of answers, and thus increase the consistency of the method, the following changes to the question have been made.

Question five revisions in italics.

With examples, explain why you think *different* scientists' observations and interpretations *of the same experiment are the same OR different (not how are interpretations different from observations).*

Question 10

"With examples, explain why you think scientific theories do not change OR how (in what ways) scientific theories may be changed."

What the question is supposed to determine (from SUSSI taxonomy):

"Scientific knowledge is both tentative and durable. Having confidence in scientific knowledge is reasonable while realizing that such knowledge may be abandoned or modified in light of new evidence or reconceptualization of prior evidence and knowledge. The history of science reveals both evolutionary and revolutionary changes."

The biggest challenge for students on this question involved the final statement. While many students indicated the importance of new evidence, it was often difficult to tease out if the change to the theory was revolutionary or evolutionary. Sometimes specific word choice (such as "edited") specifically implied evolutionary change. Other times, a specific example such as Copernican Theory implied a revolutionary change. The final rubric is shown in Table 17.

Table 17. Final rubric for question 10.

Points	*Description*
3	Theories may change a) revolutionarily or b) evolutionarily, and c) either change typically requires new evidence.
2	Two of the above (typically, "yes, they may change with new evidence").
1	One of the above (typically "yes")
0	No accurate statements

The initial inter-rater reliability was low ($\rho[24] = .36$, $p = .08$)

Question 10 proposed revisions:

Do scientific theories change? If yes – how (in what ways and to what extent) and why? If no – why not?

Question 16

The initial question read "With examples, explain the nature of and difference between scientific theories, scientific laws, and hypotheses."

The word "Hypotheses" was specifically added to the SUSSI question to reflect the Dragon Core liberal studies standards for the natural sciences at Minnesota State

University Moorhead, and seemed a natural addition. The original question only asked about laws and theories. Question 15, a Likert Scale question regarding theories and hypotheses was also added (Scientific theories are hypotheses that have been tested many times and not disproved.)

In general, students tended to explicitly answer the first half of the question ("explain the nature of") by giving definitions of one or more of the terms, but often not all 3. The "differences between" often seemed to be implied in their answers instead of stated explicitly. For example when a student stated "a law is something that cannot be changed" the rater could assume that the student believes that the other two (hypothesis and theory) may be changed. The level of acceptable assumption is an obstacle to inter-rater reliability.

What the question is supposed to determine (from SUSSI taxonomy):
"Both scientific laws and theories are subject to change. Scientific laws describe generalized relationships, observed or perceived, of natural phenomena under certain conditions. Scientific Theories are well-substantiated explanations of some aspect of the natural world. Theories do not become laws even with additional evidence; they explain laws. However, not all scientific laws have accompanying explanatory theories."

The expected student misconceptions were present (Laws do not change, theories can become laws with more testing). The final rubric for question 16 is shown in table 18.

Suggested changes to question 16 include three sub-prompts for the definition of each (theory, hypothesis, law), and a final sub-prompt explicitly about the differences between them.

Table 18. Final rubric for question 16.

Points	*Description*
3	Students need to make correct statements about each (definition), and explicitly state 1 distinct difference, and have no incorrect statements (laws explain theories, theories become laws, etc.).
2	2 correct statements and a distinct difference, 3 good definitions) and no incorrect statements or 3 + 1 correct (above) and 1 incorrect statement.
1	More correct statements than incorrect.
0	More incorrect statements than correct.

Initial inter rater reliability was again low. ($\rho[25] = .1369$, $p = .66$

Question 21

"With examples, explain how society and culture affect OR do not affect scientific research."

What the question is supposed to determine (from SUSSI taxonomy):

"Scientific knowledge aims to be general and universal. As a human endeavor, science is influenced by the society and culture in which it is practiced. Cultural values and expectations determine what and how science is conducted, interpreted, and accepted."

Table 19 shows the final rubric for question 21.

Student answers tended to focus on only one aspect, such as a culture having a taboo against researching on dead bodies, thus limiting what science is conducted or acceptable. A few specifically mentioned the church rejecting scientific findings in the past. Few students, however, addressed all three parts. Follow up interviews tended to

Table 19. Final rubric for question 21

Points	*Description*
3	Answer includes at least three of the following: That culture shapes how science is 1. conducted (methods), 2. interpreted and accepted i.e. how it is 3. "allowed" beforehand or 4. "believed" afterwards.
2	Answer includes two of the above.
1	Answer includes one of the above.
0	Provides no specific correct examples or explanations.

Initial inter-rater reliability was good ($\rho[25] = .75$, $p < .001$).

reveal that students had this knowledge when led (if the student's written answer focused on limiting what science was conducted, when specifically asked what influence culture could have after the experiment is performed or vice versa typically yielded the correct response).

Thus, more specificity to the question might be appropriate. Question 21 revised: "With examples, explain how society and culture affect OR do not affect scientific research. Think specifically about how society may or may not influence an experiment before, during, and/or after the experiment is completed."

Question 26:

"With examples, explain how and when scientists use imagination and creativity OR do not use imagination and creativity."

What the question is supposed to determine (from SUSSI taxonomy):

"Science is a blend of logic and imagination. Scientific concepts do not emerge automatically from data or from any amount of analysis alone. Inventing hypotheses or

theories to imagine how the world works and then figuring out how they can be put to the test of reality is as creative as writing poetry, composing music, or designing skyscrapers. Scientists use their imagination and creativity throughout their scientific investigations."

Many students were adamant that scientists should not be creative with their data, in the sense that scientists could not "make up" data. A small proportion thought creativity might be necessary in coming up with the problem to be researched or the hypothesis. Only a few hinted at creative/novel procedures, data analysis techniques, or conclusions. I suspect much of this difficulty stems from a lack of experience in science, where a new approach (such as graphing the results on semi-log paper) might make a relationship far more clear than the initial data might suggest. Changing the question to specifically reflect creativity at each step of the scientific method should help, in theory, to elicit answers with a greater number and variety of ways in which scientists use creativity. However, since many of these analysis techniques are of a mathematical nature, math ability could be an impediment to providing examples and experiences for the average student that would illustrate this form of creativity. Table 20 shows the final rubric for question 26.

Question 26 revised: "With examples, explain how and at what step(s) in the scientific method scientists use imagination and creativity OR do not use imagination and creativity."

Question 31

"With examples, explain whether scientists follow a single, universal scientific method OR use different types of methods."

Table 20. Final rubric for question 26.

Point	*Description*
4	Students mention that creativity is used in at least 4 of the following ways: 1. developing a research question 2. developing a hypothesis 3. developing a procedure 4. conducting analysis of the data 5. developing a conclusion.
3	Students mention 3 of the above.
2	Students mention 2 of the above.
1	Students mention 1 of the above.
0	Scientists do not use creativity.

The initial inter-rater reliability was moderate ($\rho[25] = .61$, $p = .001$).

What the question is supposed to determine (from SUSSI taxonomy):

"Scientists conduct investigations for a wide variety of reasons. Different kinds of questions suggest different kinds of scientific investigations. Different scientific domains employ different methods, core theories, and standards to advance scientific knowledge and understanding. There is no single universal step-by-step scientific method that all scientists follow. Scientists investigate research questions with prior knowledge, perseverance, and creativity. Scientific knowledge is gained in a variety of ways including observation, analysis, speculation, library investigation and experimentation. "

Student answers tended towards the idea that there was only one universal scientific method, with only three students giving any indication of how different methods of science are conducted. Even during the follow-up interviews, students who had taken both observational science classes (such as astronomy) and experimental science classes (such as chemistry) did not immediately hit on the difference until a much

Table 21 Final rubric for question 31.

Points	*Description*
3	Specific example of cases where scientific methodology would differ.
2	Vague impression that methodology is context specific.
1	The overall method is the same, but there might be a few differences, without any specifics as to how or why.
0	There is only one scientific method that all scientists universally follow.

more specific example (what essential parts of an experiment to test a drug on mice would not be appropriate to research in astronomy?). Again, student lack of experience doing science, instead of merely reading about it, is likely to blame and hard to query without providing too guiding an example (as in the interview, above). Table 21 shows the final rubric for question 31.

The initial inter-rater reliability for this question was moderate ($\rho[25] = .5058$, $p = .01$).

Possible change to question 31:

"With examples, explain whether scientists follow a single, universal scientific method OR use different types of methods. Reflect upon the labs and activities that you did in a variety of science classes."

Question 59

"What is the most important characteristic of a model?"

This question attempts to address several points the literature makes about specific misunderstandings students have of models. The rubric is roughly based on the idea that naïve, level one modelers tend to think models must be exact replicas of the target; therefore, being exactly like the target would be of highest importance. The

Table 22. Rubric for question 58.

Points	*Description*
3	The most important purpose of a scientific model is to make a hypothesis, aid in experimentation, and/or make accurate predictions.
2	The most important purpose is clear communication. Functional similarities are more important than physical similarities.
1	A scientific model should be an exact replica of reality. The purpose is to see/show.
0	No answer or none of the above.

second level indicates models are approximations, and may have different points emphasized or deemphasized for clarity. These deviations from an "exact replica" aid in communication, and indicate the model was designed for a specific purpose. There may be some indication in student answers at this point that similarity of function/behavior is important as well as physical similarity. Finally, at the third level, students express the importance of a scientific model to aid in the design and testing of experiments and/or to have great accuracy in its predictions. These two ideas go together because having the model allows for theoretical predictions, which the scientist can then try to verify experimentally. Table 22 shows the final rubric for question 58.

The initial inter-rater reliability was ($\rho[25] = .3866$, $p = .0563$)

Question 60

"Multiple models exist of the same phenomenon, such as the atom. Why?"

The intent of this question is to probe multiple models. Expected answers would indicate that when one is first taught to draw an atom, typically one includes all of the

Table 23 Final rubric for question 60.

Points	*Description*
3	Multiple models typically focus on predicting different aspects of the target's behavior. Emphasis on complexity/accuracy trade-off.
2	Focus on communicating different aspects of the target. Emphasis on making one part clearer and communication.
1	Focus is on showing, such as provides more physical detail. Vague "different ideas" statement or "points of view".
0	Not a correct statement, for example, multiple models exist to accommodate learning styles.

particles in the nucleus, and each electron in each shell. Later, Lewis Dot Structures, which look less like an atom but are quicker, easier, and more useful, are used to indicated bonding between atoms (although they tell nothing of the inner shells or nucleus). Perhaps other models would be mentioned that do a better job of conveying accurate scale, three dimensionality of atoms, or the uncertainty of the location of the electron. Unfortunately, the context chosen (atom) may not have been sufficiently familiar to the students to elicit good responses. Given the various backgrounds of these students, it is uncertain what scientific phenomenon would elicit the appropriate responses for all students, even given that they understood the reasons for multiple models. Interestingly, the "different learning styles answer" reported elsewhere was not present here. Table 23 shows the final rubric for question 60.

The initial inter-rater reliability was moderate ($\rho[25] = .58$, $p = .003$).

Question 60 revision.

Table 24. Question 61 final rubric.

Points	*Description*
3	At least 1 physical, 1 abstract, and 1 mathematical model are listed.
2	Two different types of models are listed.
1	Only 1 type of model (typically a physical model) is listed.
0	Nothing written or only non-models are listed.

"Multiple models exist of the same phenomenon, such as a map of the United States. Why?"

Question 61

"List as many science models as you can think of."

Most modeling tests and/or interviews start subjects with a similar question. It is used to determine if students can think of other models besides physical models. Table 24 shows the final rubric for question 61.

The initial inter-rater reliability was low ($\rho[24] = .37$, $p = .07$).

This question would likely remain unchanged.

Question 62

"A headline reads "Global warming model predicts that sea level will rise 2 meters by 2100 AD". What do they mean by "model" and how was this model created?"

This question was designed by the researcher to assess understanding of a mathematical model in a contextual situation. The use of numbers in the question strongly implies a model based on some sort of equation and variables, rather than a physical model.

Table 25. Final rubric for question 62.

Points	*Description*
3	The model was an equation, formula, computer program, etc. The model was created with conscious choices about what variables should influence the sea level, examining trends in those data, making assumptions about how those data are likely to change over the coming years, and then "running" the model.
2	A trend was seen in the existing data and extended. The model was built to act like the real thing.
1	Vague simulation, computer, or observing where the trend in the data is going statement.
0	None of the above.

Despite the clues, many students gave answers such as "build a small replica," implying that they had no concept of models beyond physical models.

Many did mention computers would be involved, but were vague about how conceptually (as in what variables were identified and how they were related to the sea level rise) vs. how mechanically (as in using a computer) the models were created. This might be a difficult answer to get from the students without overly leading them.

The initial inter-rater reliability was moderate ($\rho[24] = .57$, $p = .003$).

Because of the seemingly large level of initial agreement between raters, it is anticipated that this question and rubric will remain unchanged.

APPENDIX C:

INTERVIEW PROTOCOL

Students participating in the interview following the pretest will be given a code of 1, and students not participating in the pretest interview will be given a code of 0. The gains of these two groups will be compared to determine if participation in the interview seemed to be associated with greater gains.

After the Pretest Interview Protocol

1. All interviews will be audio-recorded and reviewed.
2. Students will be provided an uncorrected copy of the questions and their answers to the pretest.
3. For multiple choice questions, only questions that are answered differently than the norm of the class will be discussed. Students will be prompted to explain what they were thinking when giving that answer.
4. For each of their free response questions, the following additional prompts will be provided:

   Question 6. Ask about their science background.

   - If they have chemistry in their background, ask if they remember Charles' Law, Boyle's Law, or the Ideal Gas Law. Ask if they remember the Kinetic Molecular Theory. Ask what relationship, if any, these principles have. If students remember the laws, but not the theory, ask if the law explains why the gas behaves that way? If not, what does? If the student

does not remember anything about gas laws, ask if they remember Dalton's atomic theory, and the Law of Definite Proportions and the Law of Multiple Proportions.

- If the student has physics in their background, ask if gravity is a law or a theory and why? Most interviewees respond it is a law. Ask why gravity works. Ask if the students have ever heard of the Grand Unified Theory.

Question 11. Again, ask about their science background.

- First ask if they remember the debate about what was at the center of the solar system, and try to use that debate as the specific example of how and why a theory changed (example of a revolutionary change). If they do not remember the heliocentric/geocentric debate, ask if they remember Lamarckian Evolution and how it differs from Darwin's Theory of Evolution through natural selection (another example of a revolutionary change).
- If they have chemistry in their background, bring up Dalton's Atomic Theory. Ask if they remember the parts of it. If not, provide the parts. Ask if any parts of the theory have changed and why (example of an evolutionary change, because atoms are no longer considered indivisible or the smallest pieces of matter).
- If they have biology in their background, bring up Cell Theory. Ask if they remember the parts. If not, provide the parts. Ask if any of the parts have changed and why (example of an evolutionary change, because

viruses, for instance, are typically considered a living thing, yet are not made of cells).

- Ask students to reflect on the difference between the revolutionary examples and the evolutionary examples.

Question 12. Ask students about models that they used in classes. Prompt specifically about ecosystems (in order to generate responses of food chains, carbon cycles, and other non-physical models), math classes, economics, and geology.

Question 13. If students do not have a good answer to this question, ask them to list out the different kind of maps they may have used in a geography or history class. Would all of these maps help a person to drive from Minnesota to LA? Why or why not? Refocus student on scientific models. Ask if they remember any different models of the atom that they used in chemistry. Why were there so many different models? What did they use the models for? If they do not remember multiple models of the atom, ask about ways to represent what eats what in an ecosystem (food webs/chains/trophic pyramids). Why is it necessary to have three representations for what eats what?

Question 14. Again refocus on the multiple models of the atom and/or ecosystem to see if students will elaborate on the idea that different models convey different information, with different levels of detail.

Question 15. Prompt with the idea that two meters and 2100 AD are fairly specific quantities. What might that signify? If students give an answer

relating to graphs, computers, or equations (which is how, physically, as in using what instruments, the model was made), ask how (conceptually) the modeler knew what information to put on their graph, in their computer, or in their equation.

After the Posttest Interview Protocol.

1. Interviews will be audio-recorded and reviewed.
2. Students will be provided an uncorrected copy of the questions and their answers to the pretest and posttest.
3. After sub-scores from the pretest and posttest are compared, sub-scores showing large gains (50% raw gain) and no or negative gain will be selected. An order of questioning will then be established to focus on questions where student ideas either improved dramatically or did not improve.
4. Students will be asked to look at their pretest and posttest answers for each question in turn on the list, until the 30 minutes expires.
    - If their score changes (either increases greatly or decreases), they will be asked to explain why their answer changed. Specifically, if possible, students will be encouraged to point at the activity, etc. that they feel directly led to their change.
    - If their answer does not change and
        a. their answer was correct to begin with and no further gain is possible, this question will be skipped.
        b. their answer was incorrect to begin with, students will be prompted further that their initial answer was not the best answer available.

Given that knowledge, can they explain what the best answer is and why their answer is not the best? Since this feedback is being given after the posttest, it will not bias the data.

APPENDIX D:

MODELING ACTIVITIES

Human Population Lab

Turn in 1 packet per group, and one 1-page conclusion per person.

We are going to study some demographic data and trends around the world. Some directly relate to the environment (energy use, CO_2 production, environmental treaties), while others indirectly relate (population and income).

One thing to keep in mind when doing this activity is that the U.S. often falls at the bottom of the developed world (Europe, Japan, Canada, Australia), perhaps because there are two Americas – mainstream America and the America that is being left behind (certain minorities, isolated rural communities in Appalachia, etc.) The conditions of this second group (life expectancy, infant mortality) are often closer to that of the underdeveloped nations, and this drags down the national average to below those of other developed countries. The question becomes, then, do those other countries lack minority groups or do they just treat them more equitably and integrate them more fully?

Get into your discussion groups.

1. You will need to gather information about 9 countries. Select a country from each of the lists below, and enter it on the TOP of your Data Chart. The starred countries are suggested.

A. (New World)	B (W. Europe)	C (E. Europe)	D (AIDS Africa)
*United States	France	*Russia	Zambia
Australia	Norway	Poland	*Zimbabwe
Canada	*United Kingdom	Bulgaria	Botswana

E (Non-AIDS Africa)	F (Latin America)	G (Muslim)	H (Population Reform)
*Nigeria	Brazil	*Egypt	China
Congo	*Mexico	Iran	
Chad	Columbia	Saudi Arabia	

J (South East Asia)

*Bangladesh

Kampuchea

India

Open in a new tab the following links:

a. http://hdr.undp.org/hdr2006/pdfs/report/HDR_2006_Tables.pdf (the Human Development Report)

b. http://www.nationmaster.com/index.php (Nationmaster)

c. http://www.prb.org/pdf07/07WPDS_Eng.pdf (World Population Data Sheets, 2007)

POPULATION PREDICTIONS

2. Page 2.

a. Which of the top 10 countries (by population) is/are expected to double in population by 2050?

b. Which countries currently in the top 10 will no longer be in the top 10 by 2050? Do they have anything in common?

c. Which countries currently NOT in the top 10 will be in the top 10 by 2050? Do they have anything in common?

3. Write down the 2007 population for each nation on your sheet.

4. Write down predicted 2025 and 2050 population for each nation on your sheet.

5. Write down the % natural increase and projected % population change for each nation on your sheet.

6. Where is growth occurring? Where is it not occurring? Comments, explanations?

7. Look at pg. 4 of the World Population Data Sheet, and the information regarding birth rate and education. Do you think this is a *correlation* (birth rate goes down as education goes up, but one does not directly cause the other) *or causation* (high education causes lower birth rate)? Why? If only a correlation, explain what other variable(s) might be causing the apparent relationship between birth rate and education. If you have time, there is a whole section of the HDR (pg. 371 (89 of 110)) that is devoted to education, particularly education of women, and you could check each of your countries there to see if there was even more of a relationship. These ideas are important because some groups believe that birth rates around the world can be lowered with education. COMMENTS:

8. What is the life expectancy (total) in your countries (starts on pg. 11)? Write on chart.

9. What is the population density of each country? Write on chart. Is that country's population expected to grow or remain stable by 2050? Is there a limit to how crowded it can get? What do countries historically do when they run out of land for their people?

10. Find the infant mortality rate (deaths per 1000). Write this number down for each country on your chart. This is a good measure of the overall level of disease and parasitism (worm infestations, etc.), sanitation, nutrition, health care and living conditions.

11. Another measure might be the percentage chance not to live to 60 years of age - Pg. 295 (13 of 110) in the HDR document for highly developed countries, or percent chance to not live to age 40 - pg. 292 (10 of 110) for less developed countries. Which country in the top 25 has the highest chance to die before reaching age 60? Are you surprised? What explanation can you offer?

12. So what is the point? Population is a contributing factor to most of the environmental problems facing the world; more people produce more waste, need more food, need more energy, use more resources, etc. If the world's population continues to grow at 1.2% per year, and if each person makes 1.2% less waste and uses 1.2% less energy each year, then the world breaks even. This rate of increase works out (compounded) to a 20% cut per person in 15 years but it does NOT get us ahead, because there will be a corresponding 20% increase in people. This is assuming that no "less-developed" countries try to become industrialized, nor any "medium-developed" countries try to move up to "highly-developed" ... and highly polluting.

There are two sides to this argument. Developed countries tend to want to point the finger at less-developed countries and tell them a) the world cannot afford for your population to become developed, so you must stay less-developed and/or b) stop your population growth.

Underdeveloped countries tend to point their finger at the more-developed countries that have been polluting the air and water since the industrial revolution 200 years ago and say, "you need to cut back on pollution/development now, and then there will be an opportunity for us to develop up to a sustainable level". YOUR THOUGHTS?

Finally, looking at the life expectancy and infant mortality numbers, what happens when the less-developed countries become developed and their life expectancy increases and their infant mortality decreases? When almost every baby born lives to adulthood, and adults live to be 70 instead of 35? What will happen to population growth then?

Should population growth be checked (a moral question)? If so, how? China has succeeded with 1 child per family, but at what cost? Bangladesh tried a sterilization program with cash incentives: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=12260382&dopt=AbstractPlus, and so did India: http://query.nytimes.com/gst/fullpage.html?sec=health&res=9C0CE1D71E3DF937A25750C0A966958260. But consider what this means. Is this class warfare? Who do you see at plasma donation centers or at PRACS studies here in the Fargo Moorhead area? Is it the kids of the wealthy? Who would be likely to accept cash (it was approximately $10 U.S., which was a month's wages at the time in India) to be sterilized? The rich?

COMMENTS:

13. Write down the % Adult population with HIV/AIDS on your chart. What does this mean to a young adult living in this country? How would you feel about dating if you lived there?

14. In each of your countries, what percent of natural habitat remains (starts page 11)? Write on chart. What would you expect to happen to this number in a rapidly growing country?

How can this habitat be preserved? We try to make Brazil "save the rainforests," but people in other countries might like to see the United States put much of our farmland back into natural habitat. How does that feel? What right do they have to tell us how to use our land? What right do we have to tell Brazil how to use its land?

<u>$$ Money/GNI PPP/Standard of living$$</u>

Some people argue that when a person has to worry about where the next meal is coming from, that person will not worry about the biodegradability of the wrapper that the food came in, or worry about disposing of that wrapper properly afterwards. In other words, on Maslow's hierarchy of needs, being worried about the world environment comes pretty late. Thus, it might be productive to see where the rest of the world sits economically.

15. Write down the GNI_{ppp} per capita ($$ per person, adjusted to U.S. purchasing power) of your countries on your chart (or GDP per capita from the HDR).

16. Assume that the world goes through good times, and every country's total GNI increases by 50%. Everyone is better off, right? Not so fast. What is has happened to this country's population in this time? If it is expected to increase by more than 50%, then the number of people to spread that money around to increased faster than the money did, and that country actually gets poorer, per capita. Do you see where the phrase "the rich get richer and the poor get poorer" comes from? Which of your countries would get richer and which poorer? comments:

Can a country bring itself out of poverty while rapidly increasing in population? Why or why not?

However, the GNI_{PPP} does not tell the whole story. This is a *mean* average (add them all up and divide by the total) and thus is subject to outliers (a few very rich individuals raising the average, when the majority of the people are considerably poorer than the average would indicate). While it is harder to find median or mode incomes for each country, the Gini Index is one way to measure the spread of incomes. A Gini Index score of 0 is total equality (perfect communism?) and a Gini Index score of 1 represents 1 person having all the wealth and everyone else having nothing (perfect capitalism?). A few Gini Index scores are on pg. 3 of the World Population Data Sheets, but a complete list can be found at http://hdr.undp.org/hdr2006/statistics/indicators/147.html. Write down on chart. A better explanation of what it is can be found at http://en.wikipedia.org/wiki/Gini_coefficient.

Which country is the ONLY country among the 20 most developed in the world with a Gini Index above 20? COMMENTS?

When compared to the GDP index (http://hdr.undp.org/hdr2006/statistics/indicators/8.html) it is possible to see that most of the top 20 countries are within .05 (5%) of the top. Thus, the "average" person in the U.S. has slightly more purchasing power than other countries. However, the higher Gini Index score for the U.S. makes it less likely that there is an "average" American, just a few very wealthy and the masses of working poor. For a different perspective (http://en.wikipedia.org/wiki/Demographics_of_the_United_States scroll to the bottom), Wikipedia puts 84% of the U.S. in lower middle class or below, with only 16% above

lower middle class. On page 295 (13 of 110) of the HDR, there is a list of what percent of the people in each country live on less than $11 a day, or less than 50% of the median income. Do these numbers support the Gini Index scores for each country on your list? Comment:

How do the infant mortality rates compare to the GNI_{PPP} per capita rankings? If the GNI_{PPP} per capita goes down (in 2050) because the population is growing faster than the GDP, what would you expect to happen in 2050 to infant mortality and disease?

17. Now go to http://hdr.undp.org/hdr2006/pdfs/report/HDR_2006_Tables.pdf and look at the human development index for each of your countries. (Reports start on about page 2). Record these numbers on your chart. Then rank below.

Human Development Index

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

How do these rankings compare with what you found above for disease, per capita income, etc.?

Environmental Issues and treaties:

18. How did your country vote on the Environmental issues? About page 353 (71 of 110) of the HDR, you will see some issues regarding the environment. Key is per capita CO_2 emissions. Who is producing the most greenhouses gasses per person? (Write down CO_2 per capita and electricity consumption per capita for each country). The least? Sometimes, if you look at the GDP per unit of energy, you can see why - supply and demand. If it costs more, people use less, and vice versa. For instance, of the 1st 25 countries, Hong Kong and Greece have the most expensive energy, and also produce the least CO_2 per capita. In light of this, what can you say to the idea that raising prices/taxes on energy would decrease consumption and thus CO_2 production in the United States?

Next is ratification of 4 treaties. For instance, the U.S. did not sign the Convention on Biological Diversity. Why would they not sign something the rest of the developed world signed? How do your countries rate?

Nationmaster. Go to the Nationmaster site and find 5 more statistics about your countries, with at least one relevant stat each from Energy, Environment, and Health or Mortality. Appropriate stats would be statistics such as Municipal Waste per capita, NOx emissions, per capita nuclear energy production, etc. Inappropriate stats would be Snow Leopards (not all countries had them to begin with). Look at the list, particularly at your countries from above. What do these stats tell you?

1.

2.

3.

4.

5.

Students look at the Human Development Index (HDI) for a variety of variables and a variety of countries. The HDI is a model used to predict how good life is in a particular country. The HDI goes beyond money and looks at other factors like the status of women, life expectancy and other health issues, crime rates, environmental issues, etc. and condenses these issues into a single number. You are probably not used to this kind of a "model" yet. But let us critique the model – what do you think about the inputs the creator used to arrive at this ranking? Do you think these are valid inputs/assumptions? Are there other assumptions you would include that they did not, if you were to rank the countries on their quality of life?

Resource Lab

What did you eat this week?

Do the best that you can, and try to put each item where it makes the most sense.

Beware of serving sizes (it might help to read the box). A serving size for milk is one measuring cup, not the 32 oz. big gulp collector's glass you drank it out of which holds 4 servings. If you eat cereal out of a mixing bowl, adjust the numbers accordingly, etc. It is not necessary to do the math on this worksheet, as the Excel spreadsheet will complete all the math for you in class.

Grains/carbohydrates SERVINGS

Bread (bagels per half, buns per half, etc.) ____ slices / 8 = ____ pounds

Pasta, rice, oatmeal per cup cooked ____/8 = ____ pounds

Cereal dry (add milk below) ____/8 = ____ pounds

Candy per bar ____/8 = ____ pounds

Regular pop/Kool-Aid/beer, per 12 oz. can ____/10 = ____ pounds

(x1.6 for 20oz pop) or 1 shot

of liquor

Potatoes ____/4 = ____ pounds

Total carbohydrates ADD to get ____ pounds

1. About 7 pounds of rice (grain) would provide the energy requirements for a person for a week. Are you meeting your energy requirements for the week by grain alone? More than meeting it? How many times over or under?

Vegetables, fruits	SERVINGS	
Servings of vegetables (4 oz.)	____/4	= ____ pounds
Fruit (1 apple, orange, etc.)	____/4	= ____ pounds
Juice, wine (4-6 oz.)	____/4	= ____ pounds
Total		= ____ pounds

2. Vegetables are usually not as efficient as grains in calories provided per acre. The same is also true with fruits, only more-so. 10 pounds of vegetables might not have the same calorie value as a pound of grain. If they are more inefficient in providing calories, what do fruits and vegetables provide that grains do not?

Animal items:

Most animals are grown in intensive conditions in the U.S. That means feed lots and grain/hay instead of grazing. Even though animals like beef cattle are often grazed young, they are often put in a feed lot to "fatten up" or "finish".

Each animal has a different rate of converting feed into meat. This part will get a little tricky.

		Conversion	Waste	Lbs. Grain Equiv
Pounds of Beef.	_____	X7.5=_____	X2=	_____ _____
Pounds of Chicken	_____	X4 =_____	X1.2=	_____
Pounds of Pork	_____	X2.4 =_____	X1.3=	_____
Meat Total				_____lbs.

Do you eat more pounds of grain directly as grain or indirectly (through meat)? How many times more? Is this the best use of our resources as a planet? Remember 7 lbs. of grain would feed a person for a week. How many people could have eaten the grain that was fed to your meat animal to make the meat you ate in 1 week?

But… Meat remains the best source of many nutrients, including protein, iron, and vitamin B-12. Anemia, Kwashiorkor, etc., are also problem diseases. Comment on this trade off.

	Servings	equivalent pounds of grain
8 oz. glasses milk	_____/2 =	_____
or yogurt or ice cream etc.		
2 oz. cheese	____/2 =	_____
Dairy Total		_____lbs.

How many pounds worth of grain were fed to your animal to make the dairy products that you eat a in a week? Again, compare to the pounds of grain you eat directly. Comment.

Bonus Resources

Lbs. of Caught Fish	____ x1 = _____	x1.2=	_____**
Lbs. of Farmfish	____ x2 = _____	x1.2=	_____**
Hunted meat	____ x7.5= _____	x2 =	_____**
(Or 100%pasture on a rocky slope, swamp bottom, etc.)			
Total			_____**

If an animal has been grazing land that is unfarmable (too steep, etc.) aquatic, or set aside for conservation, eating it does not decrease the amount of food that could be grown, so it is a bonus. However, animals in these conditions grow slower so this is not

a large part of the meat available *in this country*, unless you kill it yourself. Part of sustainable agriculture that I saw in France one summer focused on "Hardy Breeds" of cattle that could be raised in areas where no other agriculture was possible. Milk was also collected in this way, as portable milkers were hauled by tractor (waste of energy??) into the mountains where the cows grazed on natural grass, instead of using feed lots. Yields are lower (it takes 4 years for the cattle to reach market weight, instead of 1, and milk was limited to 11,000 lbs. per year instead of the near 30,000lbs in feed lot conditions for Holsteins), but they are BONUS resources. Humans have little other ability to get food from untillable land.

Comment on these ideas. Could it work in America? What are the advantages, disadvantages? The French subsidized these farmers heavily ($100/head?) to make it work, but they were encouraging a sustainable agriculture. Is this a good idea in the long run (for the world)? Is it fair to the consumer? One other item I wonder about is food prices being so low in this country. Like the gas issue, doesn't that encourage waste and excess consumption? In France, pop in stores was as expensive as beer and wine in restaurants and stores and there were no free refills (at 100 to 200 calories a glass). They were A LOT thinner than we are too. Coincidence? How do we change?

To raise farm fish (salmon, shrimp) farmers catch "trash fish" and grind them up to feed the farmed fish. This is one black mark against aquaculture. We could feed more people if we ate the trash fish directly.

Some farmed fish (catfish, carp, tilapia, scallops) either eat natural vegetation or grain pellets, so are less of an impact. Thoughts?

Cash Crops. List below the items that you used this week, as often as you used them:

Ethanol (gasoline)

Coffee (cups)

Chocolate (servings)

Tea (cups/glasses)

Cotton/linen (items)

Tobacco (packs of cigarettes, etc.)

Wool/(including cashmere, mohair, etc.) (items)

Estimate total as pounds of grain lost _____ pounds

(your guess is as good as mine)

All of the above items were grown instead of grain (tobacco, coffee), were fed grain (wool), or turned grain into another product (ethanol). Many of these (coffee, tea, chocolate) are grown in developing countries where people have insufficient food. Comment on what the usage of these cash crops means to world hunger. Should "the rich" be allowed to use them when "the poor" are starving? Could this land be better used for wildlife habitat?

Go back through and *star* all of the items that had to be transported more than 100 miles to get to your house. That includes all seafood, almost all fruit and fresh vegetables, and in this area, most chicken and some beef. The average American food travels over 1000 miles!!! Think about the energy "wasted" to do that when we could eat domestic food. Comment.

Water Uses

In this exercise you will try to approximate how much water you use each week.

How many times a week do you:

1- Drink water _____ approx 8 small bottles per gallon =_____gal

2- Use water to cook _____ x 1gal/meal----------=_____gal

3- Flush the toilet_____ x 1-4gal/flush----------=_____gal

4- Take a shower _____ x 7gal/min----------=_____gal

5- Take a bath _____ x 30 gal/bath---------- =_____gal

6- Shave with water running _____ x 1gal/min---------- =_____gal

7- Brush your teeth with water running __ x 1gal/min----------=_____gal

8- Wash dishes_____ x 30gal/load---------- =_____gal

9- Wash clothes _____ x 40gal/load---------- =_____gal

10- Water the lawn _____ x 3gal/min/head---------- =_____gal

Total personal use = _____gal

6 and 7 both imply faucet, not shower.

Indirect uses of water (substitute to the closest) for 1 week:

Meat consumption _____ pounds x2530gal/lb---------- _____gal

Rice/cereal_____pounds x505gal/lb ---------- _____gal

Potatoes_____pounds x 25gal/lb---------- _____gal

Milk_____gallons x 900gal/gal--------- _____gal

Loaf of bread_____loaves x150gal/loaf---------- _____gal

Serve vegetables_____ vegetables x 125gal/veggie- _____gal

Fruit_____ pieces/ x 40gal/piece--------- _____gal

Sugar(even as candy) _____ pounds x 125gal/lb---------- _____gal

Aluminum _____ cans x 260gal/can---------- _____gal

Subtotal indirect use = _____gal

Other indirect uses: (Average American)

Producing energy ---------- ____632 gal

Mining and manufacturing ----------_____185 gal

Commercial (jobs and services) ----------_____90 gal

Subtotal indirect use = _____gal

Total personal use = _____gal

Total indirect use = _____gal

Total water used/week = _____gal

What can be done about the amount of water you use? Really look at the numbers and think here. Everyone always gets this question wrong. How much of a help is the change that you suggest in the long run? So what should you do?

How much energy is being used to produce your food? The right column shows how many units of energy are used to produce one unit of food energy by each method.

Food crop	Units used per 1 unit produced
Distant fishing (tuna, halibut)	12
Feedlot beef (commercial U.S.)	10
Feedlot dairy (commercial U.S.)	5
Grass fed beef (Organic? Buffalo? Home?)	4
Coastal fishing	2
Intensive poultry (commercial U.S.)	2
Milk from grass fed cows (maybe organic?)	1
Range Fed Beef (Organic? South only)	0.5

Intensive grain (Commercial U.S.)	0.5
Hunting, gathering	0.1
Traditional Rice culture	0.05

The units used include physical labor, gasoline, electricity, etc. to sow, harvest, transport, fumigate, and store the food in question.

Comment on the energy used to produce the food in your diet. Do you break even?

If it is so inefficient, why do we use feedlots, etc.?

Google the word Luddite, Neo-Luddite or Anarcho-primitivism (technology is bad) and discuss any merit to their ideas based on the information above.

Bioengineering, through traditional methods such as selective breeding, through high yield hybrids, to cloning and transgenic chimeras (cutting out a gene from one species and putting it into another), although much maligned, has made great progress in increasing the amount of food available. There is a huge debate about whether GMO's (genetically modified organisms) should be made or sold. With existing global hunger, and an increasing population to further stress the food supply, comment on the comparative "good" and "evil" of playing God by making salmon that grow 4 times faster, or crops with higher yield or disease resistance.

What will happen in the future? Will the world starve? Will affluent countries like the United States have to adjust to a diet of rice and soybeans? If poor countries see affluent countries with excess food, will they ever DO anything about it (war)? Without someone putting a gun to your head, would you ever change your diet to be more eco-

friendly? What about sustainable agriculture? Will the U.S. ever change from bigger, faster, more, more, more (intensive agriculture)? Will we have any land left?

Today's in-class writing assignment.

There are several ideas about conservation and environmental science that one can learn from the previous activity.

However, I would like us to go a step beyond the activity itself and look at the underlying model upon which this activity was based.

When I first encountered this activity, it was a purely paper and pencil activity. It was also somewhat less complex. I found myself questioning some of the numbers in the activity – does a faucet really run at a gallon a minute? Does your showerhead really use 7 gallons per minute? Is a pig really that much more efficient at converting feed to meat on the table than a cow is?

Being that this is now a computer model, we can question those assumptions and adjust them accordingly. For instance, the restrooms on campus have a "gpf" number on each of the toilets – how does this compare to the 2.5 gpf in the model? Do you have a low flow shower head? Do you buy free range chickens and grass fed beef? Some of these numbers might change. In fact, after meeting with a food science major we looked up feed conversion ratios and cutting % (how many pounds of carcass = how many pounds of meat) and updated the numbers. What would be the proper procedure for changing other parts of the model?

Does the fact that some of the numbers are not *exactly* right ALWAYS change the take- home message? For instance, does changing the gpf from 2.5 to 1.5 make flushing the toilet a *significantly* larger or smaller part of your typical water usage? Think about

this when someone criticizes a model of something like global warming. Yes, perhaps there are more current numbers that can be used in the model, but that one change may or may not result in a *significant* change in the final results and predictions.

Finally, are there factors that are left out? Aluminum cans are in the model, but are plastic or glass or "tin" cans for soup, etc.? Should they be? Why were they left out? Several options exist: 1) They were just left out to keep it from getting too difficult. 2) Someone figured a ratio for the amount of aluminum a person uses in a day, and estimated the amount of other materials an average person uses, then inflated the aluminum number to take everything into account. 3) Maybe they are already being counted in that average electrical usage from industry (and maybe counting aluminum separately is double counting?)

Look back over the previous day's activity.

1. List at least one part of the model (either the one that calculates your total water usage or the model of total direct and indirect grain usage) that you think you would delete. Why is this factor unnecessary/wrong? How does deleting it make the model better? Does it make it simpler? Do you think it makes it more valid/accurate?
2. List at least one factor (in either model) that the designers of the model did not take into consideration. How would adding this factor make the model more accurate? Would the increased accuracy be worth the additional effort? Can you speculate on why the creators might have left this factor out (bias, agenda, simplicity, accuracy, or they worked it in someplace else)?

3. List at least one part (probably a way that something is calculated) of either model that you think is wrong. Why do you think it is wrong? Why did the model creator put the "wrong" factor in there (bias, agenda, using an average instead of a personal number, using a number [like pop cans] to take into account other factors [like garbage in general])? Where might you go to find the "right answer"?
4. How could a model like either one of these be used to test or create a hypothesis regarding lifestyle/diet choices and food use/water use? Give a specific example.

Staple to your inventory and hand in before you leave.

A Carbon Footprint Model

Wikipedia defines carbon footprint as "a measure of the amount of carbon dioxide (CO_2) emitted through the combustion of fossil fuels … in the case of an individual or household, as part of their daily lives." This fossil fuel use could include both direct and indirect use. For instance, an electric car is advertised as having "zero emissions," because it does not burn fossil fuels and does not have an exhaust pipe. However, where does this electricity come from? If this car was used in North Dakota, over 90% of electricity is generated by burning coal. Thus, since fossil fuels were likely burned in generating the electricity used to recharge the car, from a carbon footprint standpoint, this car is not "zero emissions." On the other hand, Vermont generates about 80% of its electricity from nuclear power and most of the rest from hydroelectric plants, neither of which emit CO_2.

Task 1. (15-30 minutes)

Brainstorm a list of the factors (variables) that contribute to your carbon footprint, either directly or indirectly (for review of direct and indirect, look back at the water use activity). What activities that you do (ignoring breathing) increase the amount of CO_2 in the air? How much do they increase the amount of CO_2? For now, just put a box around each variable; the darker the box, the bigger effect you anticipate this variable having. A number of resources will be provided later to help you figure out an exact number, but just brainstorm for now.

Also brainstorm variables that might moderate some of the above carbon footprint. For instance, it takes about 0.35 kilowatt-hours of electricity (with subsequent CO_2 production depending on state) to make an aluminum can. However, if this can is

recycled, approximately 95% of this energy (and the CO_2 that goes with it) is saved the next time around. What other variables can you identify that might lesson some of the impacts listed above?

Once you have finished brainstorming alone, discuss your lists with several other students. If you receive additional variables from them, please identify them in a different color, again indicating the strength of this variable in influencing your carbon footprint by making the box heavier for those variables with a larger influence. Also share ideas about what factors may decrease (or at least moderate) your carbon footprint.

Task 2. Gather background data

NOTE: Prior to completing this activity, it is necessary to collect the following information (from your family), and the tables below will help you organize it:

- How many miles you drive in a week/month/ or year for each vehicle.
- What is the average mileage, year, make, and model for each vehicle?
- What is the average (or total) amount of energy and amount of money spent in your house for a month or year (this can be found on a utility bill).

Please write down all sources, for instance, if you heat with LP gas and wood, how many cords of wood did you burn last year and how many gallons of LP gas did you use? For those of you living in apartments, you may NOT have access to this information because it is all included in your rent. Contact me and we will consider other options like working with a partner from a house.

How many plane trips did you take? To where?

	Year	Make	Model	Miles Per Gallon (city/highway)	Miles Driven
Vehicle 1					
Vehicle 2					
Vehicle 3					
Vehicle 4					

	Amount Used Per Month	Amount Used Per Year	Dollars spent per month	Dollars spent per year	"green" source?
Electric (kWh)					
Natural gas (therms)					
LP (gallons)					
Fuel Oil (gallons)					
Wood (cords)					
Coal					

Other?					

Other factors that some models think are important:

- What foods do you commonly eat? (fruit/vegetables/starches, dairy, meat). On average, are you vegan, vegetarian, or do you eat meat?
- How much garbage do you produce per week (lbs.)? At work? At home? Figure about a pound per gallon? If you fill a 20 gallon kitchen bag a week, that is 20 lbs. If you fill one of those new 60 gallon cans for the garbage truck with the automated arm, then it is 60 lbs. If you know you pack it tighter, raise the weight it a little.
- How much waste do you recycle per week? It takes about a 24-pack of aluminum cans to make a pound. A week's worth of regional newspapers (Fargo Forum) might be 3 lbs. The Star Tribune is probably closer to 7 lbs. per week. Think about it and give your best estimate.

Task 3. Examine the models

Now examine several (at least 3) of the models below that claim to calculate your carbon footprint. No two models will ask you the same questions or calculate your carbon footprint the same way. Most of these websites will provide an explanation of the model (formula and reasoning) they used to calculate your carbon footprint. The EPA site (#7) does a great job of that, as do #1, #2, #4, and #6. These are worth a read. In my unique case, I found all models lacking. Write down the following:

a) Your carbon footprint from each site.

b) What unique questions this site asked you (or included in their model) that you should add to your brainstorming list (Task 1)?

c) Which of the factors that you felt were very important (from Task 1) did this model not seem to incorporate?

d) Were there any factors that this website lumped together or used an average for? Why would they use an average?

e) After analyzing your sites, compare and contrast. Could you say which is "better?" Which site would you use? Would it depend? On what? Why?

Sites:

1. http://www.climatecrisis.net/takeaction/carboncalculator/
2. http://www.begreennow.com/users/calculator
3. http://www.carbonfootprint.com/USA/calculator.html
4. http://www.safeclimate.net/calculator/
5. http://web.conservation.org/xp/CIWEB/programs/climatechange/carboncalculator.xml
6. http://www.nature.org/initiatives/climatechange/calculator/
7. Environmental Protection Agency

http://www.epa.gov/climatechange/emissions/ind_calculator.html

8. British Petroleum – a "Big Oil" company

http://www.bp.com/extendedsectiongenericarticle.do?categoryId=9015627&contentId=7029058

Which 3 sites you choose to compare are up to you, although here are my suggestions:

- #7 and #8 are sponsored by different types of groups than #1-#6. Does either give a different result than those from the more "green" sites?

- #6 is calculated in a very different way than any of the others, and gave me a very different number. Interesting from a comparison standpoint?
- #1, #4 and #5 seemed fairly similar to me, so it's probably not worth doing more than one from that group.
- #3 has some unique features (because it is from overseas?)

Several of these sites also allow you to click on a link to see how they did their calculations. The EPA (#7) site does a great job of that, as do #1, #2, #4, and #6. Please visit at least 2 of these calculations pages (including the EPA site, if possible).

Task 4 (Extension)

You have been provided with a variety of data, including chemical formula, balanced equations, heats of combustion/heat values, and densities for common fuels. The efficiencies of typical power plants for each type of fuel are also provided. Based on this information, create your own model in Excel for calculating a carbon footprint.

a) First you will need to determine a formula for each variable. For instance, wood is composed of carbon, hydrogen, and oxygen, primarily as polymers of sugar ($C_6H_{12}O_6$). When burned $C_6H_{12}O_6 + 6\ O_2 \rightarrow 6CO_2 + 6H_2O$. With a molecular weight of 180g/mole, or just over 5 moles per Kg, and considering a cord of wood is approximately 1500 to 2000kg,

$$\frac{1\text{ cord} \times 1750\text{ kg}}{1\text{ cord}} \times \frac{5.6\text{ mole sugar}}{1\text{ kg sugar}} \times \frac{6\text{ moles } CO_2}{1\text{ mole sugar}} \times \frac{0.044\text{kg}}{1\text{ mole } CO_2} = \frac{2587\text{ kg } CO_2}{1\text{ cord}}$$

So, my Spreadsheet would have column A (cords of wood burned) and B (CO_2 from wood). In B, the formula B3=A3*2587 would convert whatever cords were put in A2 into kg of CO_2 in B3.

	A	B	C	D	E …	Z (total CO2)
1	cords	CO_2				
2	3	=A2*2587				= B2+D2+...
3						

b) If no formula is available, then try to look for an appropriate approximation. For instance, in the above example, if the best information you can find is that 10,000,000 cords of wood was burned last year/300,000,000 people in the USA, then that gives an average of 86 kg of CO_2 per person in the U.S. from firewood. You could do these averages on a statewide or countrywide basis.

Global Warming Activity

Go to: http://phet.colorado.edu/new/simulations/index.php?cat=Light_and_Radiation then click on the "Greenhouse Effect." Then click on "Run Now," and then "Beam me down, Scotty."

As the legend states, the yellow falling "balls" are photons (or bundles) of visible light energy. This kind of light can shoot through the atmosphere pretty easily. However, if the ground, trees, etc. absorbs this light, the ground heats up using some of the energy and releases lower energy infra-red of light (the "red dots" coming up). Infra-red is the light used in the heat lamps at fast-food restaurants and the heat lamps that keep baby farm animals warm.

Infrared light tends to be reflected/absorbed and re-emitted (the red dots blink before they do) by certain molecules in the air, which are the GHGs (greenhouse gases). Notice there are a lot more of the red dots coming back down (being absorbed and re-emitted down by GHG) than there are yellow dots going up (reflected) by the ground or clouds.

Notice what the temperature does. It tends to stabilize around a certain temperature. It might randomly drift up or down a little, but stays about the same.

There are many photons; you can click the label "show all photons" and switch off a lot of the clutter (but in general, leave it on to cut down on the number of variables changing). Also at the bottom you can add/remove clouds.

Also notice on the right hand side that it defaults to current GHG conditions, but an ice age GHG setting, a 1750 A.D. setting (prior to the industrial revolution), and an adjustable setting are also available.

1. Clouds are a product of water vapor. Water vapor is a powerful greenhouse gas and is found in much higher concentrations than any other GHG. After the temperature has stabilized, click to add a few clouds.
    a. What happens to the temperature?
    b. Why? (If you are having problems, look specifically at what happens to the yellow dots falling when they hit a cloud dead on. Then answer what happens to the rising red dots. How does the combination of these two effects contribute to the overall effect?)
    c. Based upon this simulation, how would you address the climate change skeptics who claim any model that does not include water vapor (clouds) is wildly invalid?
2. Adjust the amount of greenhouse gases through the other 2 positions and write down what you observe:
    a. At 1750, a cold period in history with lower GHG concentrations
    b. Ice age, an even colder period with even lower GHG concentrations
    c. Adjustable – from "none" to "lots".

Go to:

http://www.astr.ucl.ac.be/users/matthews/jcm/jcm5/ (this will take a while) Click on "safe mode" and then "4 plots"

When I open it, it defaults to three graphs, but sometimes all four are present. Double check that you have all the graphs, which are "Fossil CO_2 Emissions – Policy" (hidden above top left), "Global Average Temperature," "Radiative Forcing – All

Contributions," and "Atmospheric CO_2 Concentrations." If you click and hold on a hidden graph, you can drag it to the empty 4th quadrant. Please do this if necessary.

1. Please look at the Radiative Forcing – All Contributions graph and list out the greenhouse gases in order of their effect (the thicker the area, the greater the effect). How does this match the notes from class?

2. Look at the Fossil CO_2 Emissions – Policy graph. This shows emissions by country/region. List out the contributors in order.

    a. Are any countries predicted to become more major contributors?

    b. What is expected to happen to the U.S. share of the total?

    c. Even though Europe and other industrialized nation's share is starting to decline, what happens to the world total? Why?

    d. Remember where the yellow and blue arrows started. As you move the cursor over them, you see one is the ability to change solar radiation, (the other is sulfur – ignore it). Does turning up the sun have the expected effects?

3. The Global Average Temperature graph shows the temperature change over the last 200 years. By moving the blue arrow, you can control what you consider to be the baseline temperature (currently 1900 A.D., but you could set it to the current year, for example).

4. You should see a giant black cross with 4 arrows on the Atmospheric CO_2 graph. You can drag this around. Please note that the x-axis is time and the y-axis is CO_2. What you are basically setting is the plateau for CO_2 – sooner or later, and at a higher or lower plateau.

a. Move the cross so it is at about the height of the light gray line (representing data points that have already been measured – you can't go "back in time" and fix points that have already been measured). What happens to the other graphs, specifically:

i. Fossil CO_2 Emission Policy. Is this even possible, explain? What would it mean about the amount of fossil fuel burned worldwide for a few years?

ii. What happens to temperature?

iii. What happens to the greenhouse gasses? Which ones level off immediately, and which ones level off gradually? This could depend on how long the gasses remain in the atmosphere or other factors.

5. There are several other features of note. First, at the top, click on maps. Then click on Regional Contributions, then Socio-economic Data. Click on each GDP, Population, and Energy. Slowly change the time. How does each variable change over time (e.g. as time goes on, China's population _____, but Africa's population _____ and then _____ starting in year _____)? Be careful, the color scheme wraps around so after a country reaching the color representing the highest possible amount of that variable will, when it increases next, go to the color representing the lowest concentration on the chart.

a. GDP:

b. Energy:

c. Population:

Do these observations match what we saw in earlier in the semester? How or how not?

6. Look at the overall plan behind this model. For instance, what lines lead into #10 global warming? Do these make sense?

7. Now it is time to look at the predictions.

Go to: http://en.wikipedia.org/wiki/Image:Global_Warming_Predictions.png

Show various global warming model predictions based on the IPCC Data Distribution Center.

a. Do all models show at least some warming?

b. Do all models agree on the exact amount of the warming?

c. Would all models support a statement that the warming would be "at least 2.0°C?" Look carefully at the chart. If all the models agree with that statement, it can be said that the model are in consensus. They cannot agree on an exact value, but they can agree on a certain minimum and maximum, or a range of acceptable values.

d. Global warming skeptics will often cite this disagreement about the exact number as proof that global warming is uncertain. The other way to look at this is that no matter how you calculate it, at least some global warming is predicted.

8. An explanation of the various scenarios can be found at:

http://en.wikipedia.org/wiki/SRES

Which of the scenarios do you feel is the most likely, based on the Population Lab earlier this semester, etc.? Support your answer. If not A2, how would using those assumptions affect the global warming predictions from #7. (above), which are mostly based on an A2 earth…

Final Modeling Project

I have included the Dragon Core competencies (see below) that you will need to use to complete the activity. I have excluded DC 1 , written communication, but you will of course be writing a short paper (~5 pages) on the results of your investigation.

The project is to take a claim in environmental science such as:

- The money spent in fertilizing (or spraying herbicides or pesticides, or irrigating) crops is not worth the return on those crops.
- The energy used to make ethanol is greater than the energy of the ethanol produced,
- The energy spent picking up curbside recycling is greater than the energy saved by processing the recycling instead of processing virgin ore,
- The gasoline saved by raising the national fuel economy by 1 mpg is more than could be recovered by drilling in the Arctic National Wildlife Refuge, etc.

First, you will make assumptions about the variables that are important to determining the answer. Then, you will research facts about those variables. Next, you will construct a mathematical relationship between the variables in Excel (if that sounds scary, it will not be by the time we get there) and make a judgment call about whether the claim is believable or not, based on your model and assumptions. You will then compare your claims with those of other people who have investigated this claim (either in class or elsewhere) and reflect on differences in a short (~3 page) paper.

This paper should

1) Define the claim to be researched, and explain why it is important

2) Explain the model

a. what variables did you choose and why?
b. what support do you have for those variables being related?
c. what confidence do you have in these numbers and why?
d. if you did not include a variable that others might expect to see, what was your rationale?

3) What conclusion can you draw from your model about the claim? How confident are you about this claim? Why?

4) Does your conclusion match that of others (in class or elsewhere) who have investigated similar claims? If not, find how their model and assumptions differ. Do you find their model more or less valid than yours? Why?

5) Use your model to make a hypothesis about what a change in policy would mean in terms of the outputs of your model. For example, if your model was paper vs. plastic bags, how many pounds of CO_2 or units of energy would be saved by mandating a switch to using only the better bag? This type of hypothesis will be considered a trivial hypothesis because it follows directly from the model, if your output predicts that a paper bag saves $.03 over a plastic bag, then if 10,000,000,000 bags are used in the United States in a year, one only needs to multiply the above numbers to find a savings.

A more interesting hypothesis would be to consider how changes in your input variables would affect the output (for instance, if your model was created 3 years ago with gas under $2/gallon, does the answer change if the price of gas goes up to $3.30/gallon?) Another alternative would be to explore what value of a variable would be necessary to reverse your decision? What is the necessary price for a barrel

of crude oil before plastic bags are the better option? At what price of landfill space does the option which produces the most garbage cease to be the cheapest option? What value must be assigned to a tree before the using of that tree as raw material becomes more expensive than leaving it in place to provide shade, provide CO_2 sequestration, prevent soil erosion, and other services? Run your model and see what your results predict, and then try to find out if anyone else has made similar prediction or observed similar results.

Rubric – final modeling project

Rubric for papers: Name:

Score_____

Points (out of 100)	Description
___/5	Topic: Unique, appropriate.
___/10	Introduction: Thesis statement, outline of supporting ideas present.
___/20	Background: Scientifically correct. Well explained.
___/30	Body paragraphs: Clear topic sentences in each paragraph, transitions between paragraphs, logical structure/**organization**, supports thesis. Unity of **focus** throughout paper.
___/10	Conclusion: Thesis restated in light of information presented in body paragraphs.
___/10	Sources: Current, credible, sufficient. Within text, correct parenthetical citations, correct use of quotes, no block quotes (a 6 page paper is too short for extended quotes). At least one source should be a primary source.
___/5	Format/Presentation: Consistent, acceptable format throughout paper. Consistent, acceptable format throughout works cited page.
___/10	**Clarity**: Spelling, punctuation, usage (fragments, run-ons, pronoun/antecedent agreement, subject/verb agreement, consistent tense), diction (contractions, slang).

Table 25. Rubric for model:

Model identifies and incorporates appropriate variables.	
Points	*Description*
3:	The most appropriate variables are included, no inappropriate variables are included.
2:	Generally correct. A small number of errors in the inclusion or exclusion of variables is permitted.
1:	A large number of errors in the inclusion or exclusion of variables is evident.
0:	No variables evident
Model integrates the variables appropriately	
Points	*Description*
3:	All important relationships are present and quantitatively correct, no inappropriate relationships are made.
2:	Generally correct. A small number of errors in the inclusion or exclusion of relationships and/or accuracy of quantitative relationships permitted.
1	A large number of errors in the inclusion or exclusion of relationships and/or accuracy of quantitative relationships is evident.
0	No relationships evident.
Model has been checked successfully against data.	
Points	*Description*
3	Model agrees with data.
2	Model agrees with some parts of the data set, but works less well with other parts of the data set.
1	Model does not fit the data well.
0	Model was not tested against the data.
Hypothesis testing	
Points	*Description*
3	Student built and tested a reasonable hypothesis from the model.
2	Student built and tested a trivial hypothesis from the model.
1	Student built and tested a hypothesis, but it was not model based.
0	No hypothesis.
Level of model	
Points	*Description*
3	Model contains postulated components or combines components in a way that is outside the typical established relationships.
2	Model contains invisible but familiar components, such as atoms, compounds, etc. for which established relationships exist.
1	Model contains only concrete physical components.
0	No model.

Dragon core competencies applicable to the final modeling project.

Taken directly from the Minnesota State University website (2006).

DC 2: CRITICAL THINKING

Goal: To develop thinkers who are able to unify factual, creative, rational, and value-sensitive modes of thought. Critical thinking will be taught and used throughout the general education curriculum in order to develop students' awareness of their own thinking and problem-solving procedures. To integrate new skills into their customary ways of thinking, students must be actively engaged in practicing thinking skills and applying them to open-ended problems.

Student Competencies: MSUM students will be able to

- Clearly define a problem and imagine and seek out a variety of possible goals, assumptions, interpretations, or perspectives which can give alternative meanings or solutions to the given situation or problem.
- Gather factual information and apply it to a given problem in a manner that is relevant, clear, comprehensive, ethical and conscious of possible bias in the information selected.
- Identify, construct, and assess arguments; generate and evaluate implications that follow from them.
- Analyze the logical connections among the facts, goals, and implicit assumptions relevant to a problem or claim.

- Recognize and articulate the value assumptions and cultural perspectives which underlie and affect decisions, interpretations, analyses, and evaluations made by ourselves and others.

DC 3: MATHEMATICAL / SYMBOLIC SYSTEMS

Goal: To increase students' knowledge about mathematical and logical modes of thinking. This will enable students to appreciate the breadth of applications of mathematics, evaluate arguments, and detect fallacious reasoning. Students will learn how to apply mathematics, logic and statistics in making decisions concerning their lives and careers.

Note: Minnesota's public higher education systems have agreed that developmental mathematics includes the first three years of a high school mathematics sequence through intermediate algebra.

Student Competencies: MSUM students will be able to

- Solve real world problems using mathematics/logical systems.
- Express mathematical/logical ideas clearly in writing.
- Organize, display, analyze information, and understand methods of data collection.
- Explain what constitutes a valid mathematical/logical argument (proof).
- Apply a variety of higher-order problem-solving and modeling strategies.

- Exhibit mastery of computational skills and the ability to make reasonable estimates.

DC 4: NATURAL SCIENCES

Goal: To improve students' understanding of natural science principles and of the methods of scientific inquiry. To instill an appreciation of the ongoing production and refinement of knowledge that is intrinsic to the scientific method. By studying the problems that engage scientists, students will comprehend the importance of science in past and current issues that societies confront. Students should be exposed to the contributions of multiple scientific disciplines.

Student Competencies: MSUM students will be able to

- Demonstrate an understanding of the scientific method and of the relationship between hypotheses and theories.
- Recognize and define problems and formulate and test hypotheses using data collected by observation or experiment. One project must develop, in greater depth, students' laboratory or field experience in the collection of data, its quantitative and graphical analysis, its interpretation, its reporting, and an appreciation of its sources of error and uncertainty.
- Exhibit knowledge of the development and contributions of major scientific theories.
- Demonstrate knowledge of the concepts, principles, problems, and perspectives of one or more specific scientific disciplines.

- Consider societal issues from natural science perspectives, making informed judgments by assessing and evaluating scientific information.

DC 10: PEOPLE AND THE ENVIRONMENT

Goal: To develop students' understanding of the concept of sustainability and the challenges we face in responding to environmental variables and resolving environmental problems. Students will examine how societies and the natural environment are intimately related. A thorough understanding of ecosystems and the ways in which different groups interact with their environments is the foundation of an environmentally literate individual.

Student Competencies: MSUM students will be able to:

- Explain the concept of sustainability.
- Identify and evaluate possible pathways to a sustainable future and demonstrate an awareness of the tradeoffs necessary to achieve a sustainable future.
- Identify the structure, function, and processes of ecosystems (ecosystems include environmental systems such as climatic, hydrologic, soils, social, and biological systems).
- Assess and analyze the environmental problems of a technological society using the framework of well-founded physical and biological principles.
- Describe the relationships between environments and socio-cultural groups, and identify how natural resource challenges are being addressed by the social, legal, economic, political, cultural, and religious systems within societies.

- Understand how socio-cultural variables affect the ways in which environments are perceived and managed, and the ways in which people or societies react to environmental challenges.

APPENDIX E:

SCORING REVISITED

This section attempts to clarify issues with the data discussed in chapter four. Following a general introduction, the data sets are handled in the same order that they were presented in Chapter Four, data analysis.

General Issues

First, some data sets were incomplete. When answers (to individual questions or complete assignments) were not given, it was impossible to determine if the lack of a correct answer related to a student's lack of ability to answer the question or if it was missing for some other reason, such as lack of time, misreading directions or questions, etc.

Second, when incorrect answers were given, it was assumed that the answer to the question was incorrect because of a lack of ability/understanding, but other issues such as lack of time, misreading the question or directions, etc. could have played a role here as well. Follow-interviews with the pretest and posttest clarify some of these concerns. Finally, the scoring of the free response items themselves did not proceed as cleanly as envisioned initially.

One problem all of these assignments did present was the influence of classmates on each other's answers. Ideally, to support the hypotheses that a particular task required a minimum Piagetian level, a hard threshold (in this case, a threshold at CTSR = 14.5, the threshold score between high and low formal operations) with only students testing at

or above that minimal threshold being able to complete the task appropriately would be preferable. On the other hand, if the task is a Piagetian task requiring high formal ability, students with CTSR scores below 14.5 should not be able to complete the task successfully. Perhaps, in a clinical interview setting this result might be expected and even achieved, but in classroom conditions, this level of control does not exist. One reason that a single CTSR score of 11 showing up in a category of scores otherwise greater than 15 (or any similar example discussed later) does not automatically dismiss the idea of a hard threshold would be that these students were free to interact with each other during all in-class assignments. Because of this interaction, a low formal student may have used a high formal student's answer, or vice-versa. There were at least two boyfriend/girlfriend pairs with substantial differences in CTSR scores who typically worked together and turned in very similar work. Furthermore, there were countless categories of friends and roommates who had differing CTSR scores who had opportunity to exchange answers. While it was tempting to throw out scores when such a ready explanation for a student testing at the concrete level (with a friend at the high formal level) gives the same high formal answer as the high formal friend, none were omitted.

The Human Population Lab

The purpose of the Human Population Lab in the class content was more central than its purpose in this study. However, as the students' first major activity, it did lend itself to an early question regarding modeling. Students were then asked, "The HDI is a model used to predict how good life is in a particular country. … You are probably not used to this kind of a 'model' yet. But let us critique the model – what do you think about

the inputs the creator used to arrive at this ranking? Do you think these are valid inputs/assumptions? Are there other assumptions you would include that they did not, if you were to rank the countries on their quality of life? Were there exceptions?"

The results were mixed, with a large number of students not answering the question in depth. In a trend that will continue with the other assignments, students indicating that they agreed with the model in question had much less to say than those who disagreed. These agreeing responses were therefore much harder to score than the disagreeing responses.

Surprisingly, the five students who did not turn in the assignment at all had by far the highest mean CTSR (19.60). Three of these students had habitual problems with missing work, but very high CTSR scores (19, 23, and 24) and another had joined the class late and so missed this early assignment (CTSR = 23).

In general, there was a fairly strong trend across all assignments that some of the most capable (highest CTSR score) students did not complete assignment.

Even though most students typed a single-spaced page or more, some students did not address the idea of the model anywhere in the assignment. Instead they tended to make off-topic comments like student 14 (CTSR = 7) who said, "I must admit I was side tracked with this lab for awhile because I kept on reading the material corresponding to the charts. All of it was very intriguing and I hope it is okay to note some of the information that I came across," and then proceeded to note information instead of answering the question asked regarding modeling. Another common thread were comments that spoke about the activity, but that did not address the question such as the comment by student 8 (CTSR = 8) who said, "I found this lab to be interesting." Student

9 (CTSR = 7) on the other hand wrote primarily about how the assignment could have been improved by making data collection and analysis in groups rather than as individuals, which again did not discuss the model.

In a recurring theme throughout this data analysis, it is difficult to know why these students did not address the question asked. Was it because they were not able to think about the model in a meaningful way (in which case these low CTSR scores would support the hypothesis), or are there other reasons that have nothing to do with modeling? Do low CTSR students have particularly poor reading comprehension and therefore do not understand what is expected? Do they have poor study skills? The answer to the last question, at least, seems to be no, as student grade point average was not significantly correlated with CTSR score (r (55) = .17, p = 0.205). Therefore, it will be considered more strongly that these students did not answer these questions for a particular reason relating to their ability to think about models in a meaningful way.

Student 17 (CTSR = 5) gave a fairly representative answer for students who had ideas on how the HDI calculation could be improved, saying,

> There are a few things that I would have added to the index if I were the one making it but non the less [*sic*] wouldn't take anything out ... One thing I would have added would have been sucide [*sic*] rates among the countries, and with that I would have added a statistic about how happy the people are that live there. I know it would be hard to calculate but it would have been interesting to find and would have tied well with the suicide numbers.

A representative answer in the *agree* category came from student 1 (CTSR = 13) who agreed with the model as is and stated why,

> When looking at the model as a whole, I think that it would be difficult to add or remove from the factors that played a part in determining the rankings of the HDI. One of the factors in the first table was the life expectancy of an average person in the country. These numbers would seem important since the higher the life expectancy, the higher the country seemed to be on the chart. This was not true to the exact age coinciding with the rank on HDI, meaning that the life expectancy was not from highest to lowest in order as the actual HDI rank was. There were countries with high life expectancy but they were low in the HDI. This is because other factors played a part in determining the rank of these countries.

Finally, a few students gave combination responses that were a mix of good and bad. Student 43 (CTSR = 18) for instance, claimed that "the model does not use any factors that involve money" (when in fact the Gross Domestic Product contributes a full third to the final score), but on the other hand, suggested adding "waste generation" to the model and successfully compared the model to the data, saying,

> When comparing Mexico to China, many of their measurements are very similar, there are a few that are better for Mexico, which results in Mexico having a higher HDI. A lower population and population density favor Mexico along with much lower CO_2 emissions. Mexico also has a higher GNIppp than China. All of this reflects in Mexico getting a better HDI.

There were several other key modeling ideas that appeared in some of the students' answers: checking the model against the data, the tradeoff between complexity and accuracy, and the purpose or bias of the creator. Each of these ideas was addressed by at least two students.

First, 12 students specifically integrated comments into their answers indicating that they had checked the HDI model against the data to verify that the HDI score did or did not match specific statistics regarding living condition between the countries in question. There was a small difference in the CTSR scores between the students who did check their model explicitly against the data (mean CTSR = 15.25) and the CTSR scores of those students who did not check their models explicitly against the data (mean CTSR = 13.67). While this did not represent a statistically significant result, the students who verified the model against the data did have higher mean CTSR scores. It could be further noted that only three of the 12 students using examples of verification had CTSR scores below 14, and none were below 9.

Although one example was given previously (student 43) looking specifically at the statistics regarding quality of life in China vs. Mexico compared to their respective HDI's, other examples would include student 29 (CTSR = 14) comparing Norway and Congo, student 38 (CTSR = 21) comparing Mexico and the United States, and student 48 (CTSR = 20) comparing England and Colombia. Student 21 (CTSR = 16) said, Realizing that the countries that had a lower income per population also had lower life expectancies, higher infant mortality rates, and some had higher HIV/AIDS percentages as well. One example of these countries is Zimbabwe. We also discovered that countries that had a higher income per population also had longer life expectancies and lower infant mortality rates. Two examples would be the United States and France.

Second, six students mentioned in their answers the idea of complexity and/or accuracy of a model being important. These students had slightly higher CTSR scores

than their peers who made no comments regarding complexity, with mean CTSR = 14.33 for those commenting on complexity and/or accuracy versus mean CTSR = 13.98 for those who did not make such a comment. Looking more closely at the data, the six students discussing complexity encompassed nearly the full range of CTSR scores (8, 9, 13, 15, 19, 20), so no CTSR threshold score for considering the accuracy/complexity aspect of models was evident. Some examples are student two (CTSR = 20) who said, "The primary reason that this model is so useful is the simplicity. The model takes many statistics and rolls them all into one general statistic." Student 30 (CTSR = 15) had similar thoughts saying, "The model is a good use [sic] due to the condensing that it does to a wide range of aspects. The use of only one figure to demonstrate a quality of life is a good idea because it does not confuse us. This allows us to just focus on the major aspect, instead of having us look at each individual number."

Third, two students commented about the creator's purpose in creating this model. While no useful conclusions can be drawn from two students, it is interesting that both students had exceptionally high CTSR scores (19). One of these two students, student 37, stated, "While the inputs the creator(s) used to arrive at the HDI rankings were very complex … you can always make it more complete and improve it. However, in theory, as long as the HDI proves to be a useful gauge of how countries are developing then it is worth taking into account." This statement was perhaps the best statement from a student regarding the idea that models are created for a specific purpose that occurred during the study.

The Human Population Lab did not pose any significant difficulty in scoring, and the above examples should provide further insight on the scoring used and the quality of representative student responses.

Resource Lab

The reflection questions for the resource lab were designed to have students specifically face the following misconceptions. First, models are static and cannot have parts deleted or changed. Once students realize models may change, the goal becomes how and why models (and by relation, theories) are changed. Second, models are neutral representations of reality, rather than a construction of the modeler that is designed with the modeler's specific purpose in mind. Third, in models more detail is always desirable. In fact there is a tradeoff in a model between accuracy and simplicity and no one model can capture every aspect of the target. It is essential that the modeler capture the most significant aspects, but too much detail could make the model too complex to be useful for its stated purpose. A fourth and final purpose was to start students thinking in terms of reasoning with models and using them to create hypotheses about how a change in lifestyle could affect the student's total water or grain use.

The four post-lab questions are repeated here:

5. List at least one factor in the model (from either the total water usage calculation or the total grain usage calculation) that you think you would delete. Why is this factor unnecessary/wrong? How does deleting it make the model better? Does it make the model simpler? Do you think it makes the model more valid/accurate?
6. List at least one factor (in either the water or grain parts of the model) that the designers of the model did not take into consideration. How would adding this

factor make the model more accurate? Would the increased accuracy be worth the additional effort? Can you speculate on why the creators might have left this factor out (bias, agenda, simplicity, accuracy, or inclusion elsewhere)?

7. List at least one part (probably a way that something is calculated) of either the water or grain aspects of the model that you think is wrong. Why do you think it is wrong? Why did the model creator put the "wrong" factor in there (bias, agenda, using an average instead of a personal number, using a number [like pop cans] to take into account other factors [like garbage in general])? Where might you go to find the "right answer"?
8. How could a model like this be used to test or create a hypothesis regarding lifestyle/diet choices and food or water use? Give a specific example.

Question one turned out to be by far the most difficult to score, with no less than 12 trends in answers emerging. In addition, two student responses also touched on the creator's purpose or bias (in addition to other categorizations).

As described in Chapter Four, Data Analysis, once these trends were established, eventually the top two categories (correct deletion and a meaningful change instead of strictly a deletion) formed one category and the other answers (that were incorrect in some respect) formed the other.

The discussion that follows gives further support to why the various emergent categories were eventually categorized as they were.

There were a very small number of students who answered this question as asked *acceptably* (only 10% of the class). The fact that each of these students had a CTSR score >14.5 (the threshold between low formal and high formal cognitive ability), as well

as the mean CTSR score of students in this category being four points above the class average points to a relationship between CTSR score and the selection of appropriate variable to delete. In order to be categorized as *acceptable*, an answer had to suggest a deletion that would not negatively affect the accuracy of the model nor detract from its purpose. Examples of this kind of deletion were the variables candy and sugar, when justified by the fact that they represented such a small part of the model that deleting them did not impact the total significantly.

The students who s*uggested changing a variable rather than deleting one* had much in common with students who suggested an acceptable deletion. Only one of the eight students in this category had a CTSR score below 15. While it is tempting to categorize these students with those who *did not answer the question asked*, their answers were conceptually different. Several students in this category suggested personalizing the indirect water used in energy production and goods and services, rather than assigning the same American average to all users. This is a suggestion that would improve the accuracy of the model. The responses in the *did not answer the question asked* category for the most part misinterpreted the question entirely, instead discussing what food they could delete from their own diet rather than what variable could be deleted from the model. Such an error may represent a misunderstanding of the idea of a model in general or the purpose of this model in particular. The students giving answers that pertained to modifying, rather than deleting, a variable did seem to perfectly understand what a model was in general, and the purpose of this model in particular. This is the distinct difference between these two categories, and was the deciding factor to group the students who gave

a well thought out modification of a variable with the students who correctly answered a deletion.

There were a variety of ways that students answered this question unacceptably.

Further analysis of the category whose answers were categorized as *did not answer the question asked* shows additional evidence that there may be a relationship between low CTSR scores and this type of response. Instead of answering which variable should be deleted from the model, they answered the question as if it asked what change to their own lifestyle they could they make or delete. Student 26 (CTSR = 13) stated, "One part of the model that I would delete would be an indirect form of water usage, my meat consumption. By only consuming 2.5 pounds of meat per week, I am indirectly using 6325 gallons of water. This is a huge amount of water, and if I were to become a vegetarian, I would not be 'wasting' this much water." In addition, since one of the common misconceptions of models is that models are static and cannot be changed, it is interesting to note that these students who gave a response that avoids changing the model itself have universally lower than average CTSR scores.

Another common incorrect response was the desire to delete the indirect aspects of the model that gave unpleasant information to the user. The indirect water used to grow food or produce goods, services, and electricity far exceeded the student's direct use of water (bathing, drinking, etc.). Likewise the amount of grain indirectly fed to animals to produce meat and dairy products far exceeded the amount of grain consumed directly for the average student. This realization that indirect use >> direct use was actually one purpose of the model, a purpose which only some students appeared to comprehended. These students who understood the purpose also understood that deleting

the unpleasant truth would make the model less able to serve this purpose. Student 51 (CTSR = 11) was an example of a student wishing to delete indirect usage variables,

> It is important to know what resources are used to produce the food that we eat … but it would be hard for the average person to change the indirect uses … so if I had to chose [*sic*] one thing that I would change it would be not to include the indirect uses in my personal total ... There are ways that people can change their direct uses of water and grain … By this change it would make the number less accurate if a person wanted to consider all the grain and water that they are consuming … With this change the model could be more realistic for change to occur at an individual level.

While in many ways, this answer shows relatively good insight (the student acknowledges that accuracy might suffer if indirect use is deleted), there is also clearly stated the idea that changes to direct water usage (flushing the toilet less, taking shorter showers) are more easily and significantly achieved than changes to indirect water usage (decreasing the goods, services, energy, and food – particularly red meat - consumed), even when facing data that the indirect usage was on average an order of magnitude (10x) larger than direct usage. Student 46 (CTSR = 18) stated, "I would definitely delete the question about soda, since I drink way too much." Student 14 (CTSR = 7) felt similarly about indirect water usage, particularly regarding meat consumption, "The indirect numbers are too high and don't really give a fair number for the average person. The model would be simpler [if indirect was deleted] since unnecessary numbers are no longer present. Finally, the model would prove to be more accurate and valid if this specific part was deleted."

In contrast, several students, student 54 (CTSR = 12), student 47 (CTSR = 10), and student three (CTSR = 8) wished to delete *essential use because essential use should not be counted against the user.* the direct usage of water. Student 54 stated, "Several factors of the model we can't control. People can't drink less water, don't flush the toilet, or don't wash their clothes. By deleting these factors the model of direct usage of water would be accurate." Student 47 stated, "To make this model more accurate, deleting a few unnecessary factors such as shower and bathroom uses might be better". Student six (CTSR = 15) wished to delete "vegetables, fruit, sugar, and aluminum" for similar reasons, "We cannot control the usage of water that these things take or have," disregarding the fact that a person may control the amount of each of these products that he or she uses. The last member of this category applied the same logic to grain; that the model should only show excess use above some baseline. For instance, if a person can subsist on seven pounds of grain per week, then this baseline should be subtracted from the total grain score, and only the grain in excess of the seven pounds should be reported by the model. However, if the purpose of the model is to allow students to compare direct and indirect use, deleting the essential use from the model would make this comparison less accurate.

Seven students belonged to the *deleted an essential part of the model because of uncertainty* category. Student 45 (CTSR = 9) seemed to be uncomfortable with any aspect of a model that involved uncertainty. This student wished to remove the entire water use category, stating, "One part of the model … that possibly could have been deleted could be the total water usage … This number is … hard to calculate … it [this deletion] would probably make it [the model] more accurate." This student also wished to

remove the indirect grain use, stating, “It is hard to know … how much the animal ate therefore it is hard to calculate correctly … You cannot really judge how big the how (*sic*) [cow?] was that you got your hamburger from.” This idea appears consistent with the misconception that models are correct copies of reality instead of useful approximations.

Question two asked students to suggest an additional variable that could be added to the model. The two *unacceptable* answers and two answers containing *both acceptable and unacceptable parts* almost exclusively suggested adding a statistic that was already present. Almost all students had *acceptable* answers such as eggs, turkey or other foodstuff not explicitly listed in the model.

Typical examples from the *acceptable* student response category follow. Student 52 (CTSR = 14) stated, “The model is missing other products we use every day that take water to produce such as plastic bottles and throw-away containers, tin cans, cardboard packaging, and fast food packaging such as Burger King wrappers and take-home Styrofoam boxes… The creator probably left this model out because it is too difficult to keep track of.” Others, such as student 48 (CTSR = 20) mentioned the omission of high-efficiency washing machines. Student 44 (CTSR = 18) suggested, “They didn’t ask about grass-fed vs. grain-fed beef.” Student 46 (CTSR = 18) suggested including the recycling process. Student 50 (CTSR = 18) mentioned the lack of tofu, but further stated, “They probably left out soy and tofu so they could make their point that eating animals is bad for the environment.” This statement, like the statement of student 52 above, shows appreciation for the purposeful creation of the model.

This model has many numbers that could easily be modified. Several of the numbers were outdated (the gallons per flush and flow rate of showers were considerably larger than current standards, for instance) and several students commented that these variables could be changed by looking at the item package or measuring the flow. Water use per meal appeared to be based on a dinner consisting of boiled vegetables and water-intensive starch preparation (boiled potatoes, pasta, or rice) which may not be appropriate to a college student in the 21st century.

On the other hand, students also had a tendency to attack the indirect usage numbers. Student 54 (CTSR = 12) stated, "I think that calculations for meat consumption, one of the factors used in the model of indirect water usage are the 'wrong' factor that suggest people should limit their consumption of meat. It is impossible to convert everyone to vegetarian, even if that would conserve water." It should be noted that the last sentence is irrelevant. Whether or not it is possible or desirable to convert everyone to vegetarianism has no bearing on the amount of water it takes to create a pound of beef for a consumer. One common difficulty for many students regarding the indirect grain and water usage through meat consumption was that student answers tended to reflect ignorance that this output was not holding them responsible for the entire water used in the cow's lifetime, but rather a proportional share of that water based on the amount of the animal actually eaten. As mentioned previously, Student 45 (CTSR = 9) stated, "You cannot really judge how big the how (sic) was that you got your hamburger from." While the instructor/researcher intervened during the activity to clarify this concept for the class when this issue became apparent, this misunderstanding persisted.

Student 55 (CTSR = 4) stated that "I don't have enough information to judge a model". Perhaps even after two activities exploring models, the student still maintained the static misconception of models. This was the only comment of this kind, but it seems to support the idea of models being correct (and beyond reproach) and static rather than useful approximations.

Since the grain portion of the model dealt exclusively with diet, and the water portion also dealt to some extent with diet, almost all student hypotheses mentioned the effect of a change in diet on resources used. Examples of acceptable and unacceptable answers are given below.

Student 44 gave an example of an *acceptable* answer when he said (CTSR = 18) said, "A great example of how to use these models as a test would be a meat inclusive diet vs. a vegan diet. It would be easy to enter (just leave meat and dairy blank on one) and would likely produce very clear results." This was by far the most typical form of acceptable answer given.

On the other hand, some students did not use the model to form a good hypothesis. The most common *unacceptable* hypothesis by far was that this model, because it dealt with food, could be used to make hypotheses regarding nutrition. Student 12 (CTSR = 9) was such a student, saying, "Athletic trainers and sports programs could use this chart to record what athletes are consuming. They could then look at exactly what someone is putting into their system and try to either lower the consumption of certain foods or tell the athlete that he or she needs to eat more of a certain food/category." Certainly a trainer could do this, but as the final result would be how much grain or water was used directly and indirectly and not whether the athlete

consumed appropriate nutrients for their health and or specific sport, this use would be an inappropriate use of the model.

In addition to the specific post-lab questions, there were a number of miscellaneous observations for the resource lab. Some students showed a very clear understanding of the purposeful creation of models. Student 44 (CTSR = 18) felt the model did not fit dormitory lifestyles as well as house-dwelling lifestyles. Student 44 suggested, "A model should, in my opinion, be very carefully tailored to the population that will be using it … If you did substitute a question that addressed dorm water use or something like that, it probably would make the model more accurate." Student 43 (CTSR = 18) stated, "They may have left this [water use by different types of animal] out because they are biased in trying to prove that meat uses the most water to produce."

Several students remarked on the complexity/accuracy tradeoff. There did not seem to be a relationship in the quality of answers between high and low cognitive ability students. Student 51 (CTSR = 11) said, "The creators of the model may have left out these factors to make the model simpler. When increasing the complexity of a model it might discourage people from using it."

In conclusion, each question contributed insight to the relationship between modeling and cognitive ability. The specific examples provided seem to support the quantitative analysis.

Carbon Footprint Activity

Of the four daily assignments, the Carbon Footprint Activity seemed to engage the students in learning about models better than the other three. As the activity and questions are primarily about modeling, and the questions are quite leading, this result is

not surprising. A majority of students answered at least part of each question correctly, and many students demonstrated understanding of the tradeoff between complexity and accuracy and the suitability of a model to an audience and/or particular purpose. There were far fewer answers suggesting that a part of the model should be removed because the student did not like the output of the model, when compared to the answers in the Resource Lab. Overall, this activity showed that most students had, in a structured environment, grasped many of the basics beliefs about models that separated someone with a level one knowledge of models from someone with a level two or three knowledge of models. The post-lab questions are repeated below:

1. What was the range of your results (low to high)? Why do you think there was such a range? What does that mean about these models? What does that mean about your carbon footprint? With such a large range, how can we use these models appropriately? Were your results in line with others who used models from similar sites (site #2 seems to always be low, or site #8 always seems to be high for instance?)

2. Accuracy/completeness versus complexity. One reason for multiple models of the same phenomenon is that certain models are more appropriate for a deeper understanding, where more accuracy is needed, and thus more complexity is required. Other times, a quick "ballpark" estimate might be appropriate. For each of the parts below, do you think the listed aspect made the model more accurate? Was the change in accuracy appropriate given the change in complexity from adding/removing that variable?

2.a. What unique questions did each site ask you (or include in their model) that you did not have in your brainstorming list (Task 1)?

2.b. Which of the factors that you felt were very important (from Task 1) did this model not seem to incorporate?

2.c. Were there any factors that this website “lumped together” or used an average for? Why would they use an average?

3. After analyzing your sites and their models, compare and contrast. Could you say which is “better?” Which site would you use? I would say "it depends." Take AT LEAST 2 of the sites and say how and why you would use one in a particular setting, but another in a different setting.

Question one was scored as follows. A score of -1 indicates that nothing was correct and one part was incorrect.

Student seven (CTSR = 9) gave this answer to question one. "The more questions asked, the larger my carbon footprint is. By asking more questions, the model can use that I make [sic] a more accurate carbon footprint, even if it is a larger number." Two students gave responses of this nature. Since some questions (such as whether or not the user’s appliances are Energy Star) actually lowered the carbon footprint score, this answer is incorrect. In general, even in a model that did not have questions that subtracted from the carbon footprint (as with the Energy Star example) a good, but simple, model should arrive at roughly the same result as a good, but more detailed, model assuming both models included the same major categories.

Two students, student 19 and 25 (CTSR 11 and 12, respectively) instead focused on the idea that there could not be multiple valid models. For example, student 19 stated “I think that these models need to have common questions that are asked in order to give more accurate readings. The large variation in numbers makes me think that there really

isn't an accurate reading for a carbon footprint." This statement reflects a key misconception that students have about models in general and about multiple models in particular, specifically that there only could be one valid model for a given phenomenon. One student, student 12 (CTSR = 9), stated that the differences in model outputs were due to monthly vs. yearly use. It is somewhat unclear what the student meant by this answer, but it is incorrect. As long as the student read the directions on the model and did not put monthly uses into a site that asked for yearly use, this monthly vs. yearly use should not have been an issue. Likewise, perhaps the student saw 1 ton/month as a different answer from 12 tons/year, but that is not a problem with the model.

Some students scored a zero as their answer had no parts correct nor explicitly incorrect. For example, student 24 (CTSR = 8), wrote a long paragraph talking about the differing results stemming from the particular bias of the creator. However, nowhere in this quote does he talk about the information used in the model nor how to best use the model, although it would have been easy for the student to mention in this paragraph that variable selection was how the bias was achieved.

> I think the differences between websites is dependent upon what message it is that the sponsors of the website want to express. For instance if a website is created by companies that support the use of coal than they are going to want to make carbon emissions appear to be minimal. If a website is created by people who want to protect the environment they will have factors that make the numbers appear much higher. What that tells me about my carbon footprint is that it probably isn't entirely accurate. I think in order for these models to be used

> appropriately and to accurately represent a person's carbon emissions it would be important to make a uniform formula that isn't biased.

Many students received a score of one. As stated in Chapter Four, almost every one of these students correctly identified why how the models arrived at different conclusions, but most either stopped their answer there without proceeding to talk at all about how to best use their model or answered that part. Various answers involving averages were suggested. Student six (CTSR = 15) suggested the mean, saying,

> I believe there was such a range because each site I used asked for different information ... The reason why they range so much in numbers is because some get really specific whereas the others just go off of general information. The reason for asking different things is the sites that ask more will be able to give you a more accurate footprint where as the ones that are just asking you the broad general questions will give your broad general footprint. With a large range of numbers we can average out all of them to get your average carbon footprint.

While the mean was the most common average suggested, other measures of central tendency were also tried. Student five (CTSR = 14) suggested the mode instead. "It may be best to try as many models as you can to see which results appear the most prevalent. " Student 21 (CTSR = 16) suggested the median, "We can use these models appropriately with a such a large range is to find the median and use that." Other students receiving a score of one simply left off how to deal with the varying models.

Several students received a score of two. Students receiving a score of two needed to have both halves explicitly correct. Student 22 (CTSR = 17) said,

> I think there was such a large range because of the way the results were calculated. There were also some different factors taken into account. These models were probably designed to illustrated different aspects of a carbon footprint. My carbon footprint could be anywhere in this range of numbers. We can use these models appropriately by finding out what they are specifically designed to show and using them accordingly.

Student 18 (CTSR = 19) had similar thoughts,

> The range most likely comes from the models the different sites have created to calculate the total tons produced yearly. Some sites will include more things such as number 2, which chose to include the types of food you eat and then subtracted carbon for recycling and other things. These actions obviously make the carbon footprint go up or down, so depending on the motives of the organization, they can include certain qualities to manipulate the data to make it tell you what they want it to. We can use these models appropriately by taking them for what they are. They are not perfect, but no model can really be.

It can be noted that both of these students specifically address the purpose or bias that might be inherent in the model. Furthermore, the second quote supports the idea that no model is a perfect representation of a phenomenon, which shows that this student does not suffer from this key misconception.

In conclusion, it can be seen that the carbon footprint activity provides insight into how these students viewed multiple models. Furthermore, there were specific cases of

misconceptions (such as there can be only one correct model) present among some students.

The following quotes from students give specific insight to representative answers from each category.

Student eight (CTSR = 8,score of zero) answered the following, "It's hard to say which one question didn't seem to incorporate. I think they all did. On website #6, my result was shown in a pie chart. This showed that they didn't lump any factors together." Here it appears this student is confusing the word *average* with the word *total*. Because the result was listed in a pie chart that showed each sub-score relating to the carbon footprint due to the home, transportation, lifestyle, etc., instead of as a single number, this student assumed no average value was used. Furthermore, this student seemed to be looking for a question that did not fit into the model, rather than a variable that was either in the student's brainstorming list from task one but not in the model, or vice versa.

Student two (CTSR = 20, score of one) gave good answers regarding variables present in the model but not in the brainstorming list and vice versa.

> One thing I never thought of, which didn't even really affect my carbon footprint was airplane flights. I was almost surprised they asked me as I never thought of it. A big focus point on my brainstorming sheet was food. I took into effect the respiration of the animals from which I eat meat, as well as the absorption of CO2 by plants, and the carbon emitted by the trucks hauling the food around the country. I know at least one of the models I used didn't even use food as a determinant.

This type of answer was very typical of the answers that received a score of one. What was present was good, and it is not really possible to know if the student could have answered the part of the question regarding averages, and just did not bother to do so, or was cognitively unable to answer the question.

Questions receiving a score of two gave complete answers that were occasionally exceptional. Student 44 (CTSR = 18) gave an excellent answer regarding the use of averages and the tradeoff of complexity and accuracy. After noting the differences between the brainstorming list (no motorcycles, no burning wood) and the models (no local foods correction, only one of the three had garbage output) the student addressed averages.

> To me there were too many averages used. I especially disliked the ranges that you had to enter at the Conservation International Site. I couldn't even enter an exact for fuel economy on my car. I had to use a range of 15 to 20 miles per gallon. Even worse was the miles drive [*sic*], you could only go in FIVE THOUSAND! mile increments … I am guessing they use averages to save people the time and effort of looking up exacts … However, it seems like they might sacrifice accuracy pretty heavily by going that route.

This answer took the opportunity to really explore the relationship between accuracy and complexity. For one political science student, student 37 (CTSR = 19), the proverbial light bulb went on and the student made a very successful, but too elaborate to reproduce here, analogy between the various political polls and multiple models, and accurate predictions as the ultimate arbiter of a successful model.

In conclusion, almost all students demonstrated they were capable of identifying variables that could be added, and this question did not pose a great deal of difficulty in scoring.

Examples of successful and unsuccessful answers for question three on multiple models are presented below. Student 35 (CTSR = 13) gave a solid answer to question three which also hinted at ho w these different models might be used to form hypotheses on particular lifestyle changes,

> The EPA personal emissions calculator would be an excellent site to see what your direct carbon footprints add up and what you can do to reduce them. For example buying a new car or making the decision to invest in instillation [*sic*, insulation?] for heating for your home and what to heat it with. In contrast the Carbon Footprint Calculator website took more indirect and detailed factors into account such as food preferences and buying habits of food and clothes. This will show the impact on the "behind the scenes" everyday decisions that we all make and how this affects our carbon footprint.

This student obviously has a solid grasp on the use of multiple models.

Student 27 (CTSR = 13, score of zero) stated,

> I would say that they all do a good job of compiling information. They all take into account usage of cars. That's about it. They all have their own curveballs. #3 and #7 take into account recycling. #1 and #3 take into account airline travel. #3 was the only one to account for motorcycles and public transit, and was all together more detailed which may account for the highest Carbon Footprint rating. It also took large liberties with how much of

> our purchases have “little packaging” and things that are “nicely packaged” as well as the “standard range of financial services.”

While lengthy, this answer does not explicitly explore how any of the models could be used, particularly to form a hypothesis.

In conclusion, it can again be stated that almost all students were successful on question three, and thus can be stated to have a solid grasp of how multiple models can be used.

Question four concerned hypothesis formation. There were many good hypotheses formed. Student one (CTSR = 13) stated, “By changing my lifestyle choices to things such as being a vegan, buying all second hand clothing and electronics, and only buying food grown locally, my lifestyle CO2 output was dropped from around 3 tons to .4 tons.” Student 30, (CTSR = 15) said, “In the website it gives you a stock starting point related to your state and then from the questions it adds or subtracts amounts to your carbon footprint. This helps show you where you make a difference and how each one can possibly affect your footprint." While not necessarily the explicit hypothesis testing seen in the resource lab, these students are looking at the cause and effect of changing an input variable and seeing how it affects the output variable.

Two other trends emerged from the study through further analysis. *Complexity vs. accuracy* was explicitly assessed in question one, but some students who did not address this aspect of modeling in question one addressed it elsewhere. For instance, Student 38 (CTSR = 21) stated, “A lot of the household impacts (such as how often you recycle and unplug unused appliances) were used as an average. An average is useful because it decreases the complexity and adds some accuracy to the overall model." Other

examples relating to the complexity vs. accuracy tradeoff have already been explored in question two.

Although not asked explicitly in the activity, all three questions could result in answers that addressed bias or purpose of the creator or intended use of the model. Examples of student answers from both high and low CTSR levels show that students were considering issues of *Bias/Purpose/Use*.

Student 28 (CTSR = 19) specifically addressed the usefulness of a model for the specific circumstances at home, saying, "I don't think this site was very efficient for apartments where I live in because it asks about do I use energy star appliances, heat and cool efficiently and usage of hot water efficiently. I have little or no say in what appliances are used and how our apartment is heated and cooled and same with the hot water." This answer by student one (CTSR = 13) covers many aspects of modeling "By keeping my car and being a vegetarian and buying secondhand appliances and clothing (etc.), my emissions still went down 4 tons. This could imply that the people who made the site are "hippies" and the numbers make one want to be a vegetarian, not own a car, etc. … I noticed that my annual output did not change in the last model even if I changed the type of heating in my house." Not only does this student address the bias/purpose of the creator, but the student also explicitly shows how they reasoned with the model.

Students of various cognitive abilities seemed to have been able to make statements regarding bias, with perhaps the best statement in the class coming from a low formal student.

Global Warming Activity.

There were complications in interpretation of the results of the Global Warming Activity. The post-lab questions are repeated below:

7.d. Global warming skeptics will often cite this disagreement about the exact number [amount of temperature increase predicted] as proof that global warming is uncertain. The other way to look at this is that no matter how you calculate it, at least some global warming is predicted. **Comment!**

8. An explanation of the various scenarios can be found at: http://en.wikipedia.org/wiki/Special_Report_on_Emissions_Scenarios . Again, feel free to go to the original IPCC report…

Which of the scenarios do you feel is the most likely, based on the Human Population Lab earlier this semester, etc.? Support your answer. If not scenario A2, how would using those assumptions affect the global warming predictions from #7. (above), which are mostly based on a scenario A2 earth…

Question seven was plagued by virtually all (over 70%) students giving the same answer, thus giving little discrimination ability to this question. This question was straightforward to score.

Student 35 (CTSR = 13) gave the following representative *incorre*ct answer, "Skeptics are using the differing range that the models predict as uncertainty. If people who believe in global warming can't agree or believe each other, what reason should people who don't believe in global warming have any reason to trust them?"

Student 60 (CTSR = 23) and student 56 (CTSR = 13) gave the following representative *correct* answers "The fact that these models all disagree about the exact

centigrade is insubstantial next to the fact that they all show warming," and "If a few of these models showed no increase in temperature, then I could understand the skepticism," respectively.

Five students did not turn in an assignment, but their mean CTSR (14.4) did not differ significantly from the students who answered the question correctly.

Eight students gave an answer that did not clearly agree or disagree with the statement. These students' mean CTSR score of 12.25 was between those of the students who agreed and those who disagreed. Student eight (CTSR = 8) had an answer typical of these students, "There is no exact answer since this hasn't happened yet. We need to wait to see what happens in the future."

Question eight, particularly in fall semester, suffered from two issues. First a large percentage of students made a content error (that the population growth predicted from readings earlier in the class was smaller than the population growth assumed in these models). This failure to recognize that the models were using an assumption that the class has a reason to believe was incorrect prevented them from choosing the answer A2 was not the best scenario. Without selecting a different scenario, it was impossible to answer the follow up question that would display the student's ability to reason with models and describe how changing assumptions in a model should change the output.

There was an apparent difference between summer and fall semester results on question eight, as only two students in the summer supported A2 (and no student who supported it explained why), whereas 10 students in the fall supported the A2 scenario and explained why they thought it was most logical. It was not clear why more students in fall supported A2, but it could pertain to the more compressed timeframe (five weeks)

of summer semester, which may have allowed students to more readily access ideas from earlier in the semester (only weeks old) as compared to the same information in the fall semester (months old).

Question eight revealed six categories of answers. Sample answers from each of the categories are discussed in the following sections.

Students who *disagreed, but did not state how the prediction would be affected,* had answers such as student 39 (CTSR = 15) who stated, "For me, the most likely scenario is A1, more specifically A1B which emphasizes on all energy sources. This is because we learned/read that global population is predicted to be reached 9 billion by 2050 and then a decrease as well as rapid economic growth of the world," or student 21 (CTSR = 15) who stated, "I think that B1 is more reasonable because the criteria seems it would be at a steady rate with no drastic changes. A2 shows that population is continuously growing for all countries and number seven shows that it is not." Thus, these students were able to complete an acceptable analysis; picking up from previous readings that world population would not increase forever, and were able to see that this idea matched better with A1 than with A2. However, for whatever reason, these students did not complete the second half of the question to conclude that therefore the predicted temperature increases of the models may be too large. One response placed in this category that was slightly different was that of student 45 (CTSR = 9) who stated, "The scenario that would be most likely to happen based on the Population Lab earlier this semester would be the A1 scenario… Because there is an increase in these things there is a chance that the CO2 levels will rise especially if there is still fossil fuels being used." While at first it appears that this student is fully answering the question with the last

statement, on closer inspection it appears that the student is only thinking about the effects of scenario A1 ("CO_2 will rise"), not on how the change from A2 to A1 will result in a change in the prediction, such as "CO_2 will rise *less*." Overall, this category easily contrasts with the category that gave the best answers, disagreed, and correctly stated how prediction would be affected.

The *disagreed, and correctly stated how prediction would be affected* category gave the expected correct answer. Student 35 (CTSR = 13) gave such an answer, stating that A1 was the more likely scenario and,

> The predictions would not show as great an increase in global warming as the graph on #7 shows as there would not be an ever increasing greater number of people continuing to be born. However, a big part of how much of an increase in temp. would depend greatly on what kind of energy was used, with an emphasis on fossil fuels or an emphasis on non-fossil fuels.

Student 38 (CTSR = 21) was even more elaborate, writing,

> Based on the population lab earlier this semester, it seemed that the world would more closely represent an A1 world, where the population will increase to 9 billion, and then decrease slightly. I wouldn't think that the world would have a continually increasing population like in an A2 world. It would also appear that the global economy will also increase steadily and our world will be more "convergent" as the model says. With these factors, it would seem reasonable to assume that our future may be an A1 world as opposed to an A2 world. With this information, the environmental models may be a bit askew from the actual events that may occur. With a decreasing population, the amount of pollution will

> decrease, so when the year 2100 comes around, global warming may become less of a problem based on this. Although it does rapidly increase, there may be an abundance of CO2 and other forms of pollutants during the first 50 years, but with the development of clean technology for energy, some of these pollutants may no longer be a variable in global warming models.

There were also a few responses placed in this category where students agreed with the A2 scenario, but hypothesized to some extent about what would happen if the other scenario came to pass instead, thus still demonstrating their hypothetical reasoning. Student 58 (CTSR = 23) captured this position when writing,

> I think that it is more like A2 … A larger population would most likely use more resources and energy and also cause a further increase in CO_2. Or what if we suddenly convert to more energy friendly fuel sources instead of fossil fuels this could also have an impact, greatly reducing CO_2 and other GHG … the sooner we convert to a greater degree to an non-fossil fuel source for driving, the more able we'll be at reducing CO_2 emission.

Student 20 (CTSR = 15) stated, "Scenario B1 is more ideal, but the cynic in me says that without serious and radical change A1 is more likely. Currently we are still too dependent on fossil fuels without one solid answer to turn to." While not particularly explicit, it was assumed by the word "ideal" that the student meant a world with less pollution and less global warming, which would in fact be the result of the B1 scenario. Similarly, student 28 (CTSR = 19) agreed that A2 was the most likely, but wrote, "If it can become a A1 or B1 world then I think we have a chance at a better future," again,

starting with the word “better,” the above logic chain results in an implied hypothesis of less global warming than predicted in A2.

The responses categorized as a*greed, but said nothing further or answer had logic flaw* was mostly comprised of answers such as that of student 59 (CTSR = 18) whose entire vague, brief answer consisted of, “I think that A2 seems the most likely in our current world. Unless there are drastic change [*sic*] in the way people around the world interact this is probably how this are going to end up.”

Finally, the *unable to be clearly scored* answers included answers such as student 10’s (CTSR = 10), “I have no clue” and student 27’s (CTSR = 13), "I think that they are both possible." Since there were, depending on the source, approximately eight scenarios, it was not clear what was meant by “both.”

In conclusion, the global warming lab questions were not as cleanly scored as questions from the other three modeling activities. However, the above examples illustrate the kinds of answers that were placed in each category.

Final Project

Initial variables submitted

These first tentative steps of their final modeling project provided unique insight into their understanding of modeling because, unlike later submissions, this first draft had not been influenced by feedback from the instructor. Unfortunately this initial probe into student understanding of the variables relevant to their model was more incomplete than other sources of data (assignments in this class indicating draft status typically had lower completion rates). This assignment revealed trends in variable selection related to CSTR

score, including the use of specific examples instead of variables, inclusion of irrelevant variables, and omission of relevant variables.

From this first list of variables came several pieces of evidence linking Piagetian reasoning and modeling. The first of these links concerns an approach where variables are fully utilized to represent an infinite number of possible combinations versus an approach where a finite number of combinations of variables are listed explicitly, typically in the form of a specific object. The first approach is more consistent with formal operational reasoning, whereas the second represents concrete thinking. During the fall sections of the class, where this second approach appeared to be more common, of the 38 students submitting a list of variables, eight took a very concrete approach (as compared to one out of 15 during the summer). Looking ahead to the final project, where students are asked to manipulate the model to create and test hypotheses, it is obvious that a model with variables will allow the modeler much greater freedom and success as compared to a model with concrete examples instead of variables. Several specific examples will illustrate this point.

Student 26 (CTSR total 11), for instance, compared emissions from American cars to foreign cars. The student submitted a list of the top 10 cars in miles per gallon (MPG), a list of four foreign car manufacturers, a list of four American manufacturers, a list of specific MPG's for 11 specific foreign vehicles and 13 specific American vehicles, and finally an average MPG for foreign cars and American cars. While these lists filled the student's page, the single variable MPG by itself encompasses all of the specific examples listed above as well as an infinite number of other possibilities. It is obvious that the student understands that the variable MPG is important in the model, but it is

equally apparent that the student has not yet grasped how a variable could be used appropriately. Another example is student 35 (CTSR = 13) using specific vehicles instead of the variable MPG and specific trips instead of the variable distance.

Compare the above example to that of Student 13 (CTSR = 21) who lists the following variables for comparison of the environmental impact of a real and an artificial Christmas tree:

- "How long do you keep a tree for?"
- "Disposal of both types"
- "How long does it take to grow a real one?"
- "How far are the real trees/fake trees shipped?"
- "Carbon footprint for each tree?"
- "Water use for real tree?"
- "Real tree changes CO_2"

All of the above represent variables necessary to answer the question. Although a few may overlap ("Real tree changes CO_2" might be part of the "Carbon footprint for each tree"), all are relevant variables, and concrete examples are not given.

Another conceptual problem was the inclusion of variables that on the surface appeared to be relevant to the situation, but in reality were not. To take the example of student 26 (CTSR 11) again, the variable of fuel tank size was listed as an important one necessary to build a model to compare emissions from American and foreign cars. Fuel tank size has appeal in that it is a concrete object, and fuel tank size is a useful variable that affects the range of a vehicle and could be used indirectly to calculate miles per gallon if a total distance driven on a full tank was known as well. However, when

comparing the emissions of two cars over 12,000 miles of driving in a year, it matters little if the vehicle fills its 20-gallon tank 20 times or fills a 10-gallon tank 40 times. What does matter is that a total of 400 gallons was burned or nearly 8000 pounds of CO_2 were emitted, and these numbers can be determined simply by knowing certain ratios such as MPG or emissions per mile.

Students often listed irrelevant variables related to current usage statistics. For example, student 3 (CTSR = 8), when comparing the relative environmental impact of using a tire as tire-derived fuel versus recycling a tire, listed "How many tires are used per year in the U.S.?" and "How many recycling or burning of tire plants are in the U.S.?" Neither of these variables can directly be used to calculate which is better for the environment. Although these variables could potentially be used as part of a longer calculation to find emissions per tire, as the other variables that would be needed (total emissions from tire-derived fuel plants and percent of tires converted to fuel, for instance) were not included in the list, it does not appear that this was what the student was thinking. Again, these irrelevant variables appear to be more concrete in nature than the more relevant and abstract variables.

A final conceptual problem that students demonstrated is the reverse of the inclusion of irrelevant variables, namely, the omission of variables essential to creating a good model. For instance, using student 26 (CTSR 11/24) once more, variables such as kilograms of carbon dioxide per gallon of fuel burned and grams of particulate matter per mile traveled are needed to determine the total amount of air pollution created during a given amount of driving. These variables that were omitted have some common characteristics that are the opposites of the included, but inappropriate variables described

above. First, as can be seen in this example, the variables omitted are not concrete objects, but rather mathematical relationships between two quantities. Second, these relationships are often ratios.

Finally, it should be noted that scoring these lists of variables proved difficult for several other reasons, not merely because of the number of un-submitted assignments. First, some variables were too vague to score. For instance, student 12 (CTSR = 9), when comparing organic to traditional agriculture, wrote, "indirect human affects [*sic*] from using chemicals on crops." It appears that the student knows that the pesticides and other chemicals are important variables. However, it does not appear that the student knows how to quantify this into ideas like "what is the safe level of nitrate in drinking water," "what percent of applied nitrate leaches from a field," "what is the environmental cost of reducing nitrate from a contaminated water supply to a safe level." Second, students submitted anywhere from three to 75 variables. Thus, one student may have both more relevant variables and more irrelevant variables than another student. Who deserves the better score? The student with more relevant variables? The student with fewer irrelevant variables? In the end at the ratio of relevant to irrelevant variables was used with variables that were too vague to be scored ignored.

The final spreadsheet project

When first envisioned, the final modeling project was seen as the final instructional tool before the posttest assessment. However, it became apparent that this project provided a central assessment in its own right, as the only assessment of the students' ability to build a model from scratch, which was sorely lacking from the SUMS pretest and posttest.

The dissertation proposal gave a rubric that assesses students on the variable selection in the model, how these variables are integrated into the model, the level (concrete, formal, or post-formal) of the model, whether or not the model was checked against data, and the quality of the hypothesis that the student formed and/or tested with the model. After examining the data, the following three additional criteria were added: were equations and variables used (in other words, was the model static or could variables within it be changed, thus changing the output), were ratios used, and finally, how was the answer expressed. With regards to the last category, was the better option merely stated (A is better than B), was the difference stated (A is X units larger or smaller than B), or was a ratio stated (A is X times or X percent larger or smaller than B)? Since these last three categories were not explicitly in the rubric (although the first two are implied), these areas are only being used to shed some additional light on the thought processes of the student.

Representative examples from each rubric score are presented below.

While details of appropriate and inappropriate *variable selections* were detailed previously in the section analyzing students' initial variable lists, it is worth repeating briefly here.

Examples of students scoring a *one*. Student 15 (CTSR = 11) focused on tank size of a vehicle instead of emissions per mile or emissions per gallon when comparing diesel and gasoline vehicles. Student 17 (CTSR = 5) was not able to resolve electricity versus wood in home heating to any kind of common unit (such as amount of carbon dioxide emitted or total dollar value of emissions). Student 19 (CTSR = 11) compared the environmental impact of two industries (recycling aluminum cans versus virgin

aluminum cans) without ever taking into account the number of units of each produced. Many of these examples seem to relate to a fixation on raw numbers rather than more appropriate ratios.

Example of a student scoring a *two*. Student 32 (CTSR = 14) had many variables identified for both a compact fluorescent light and also for a regular incandescent light, including the mass of each part (such as the glass bulb, metal socket, tungsten filament, and mercury vapor) and emissions from a power plant (such as carbon dioxide, sulfur dioxide, and mercury). However, the student used an incorrect value for the environmental cost of the sulfur dioxide and did not include a column for the environmental cost of mercury emissions. Likewise, without embodied energy values to calculate the environmental cost of the metal, glass, etc. during bulb construction, these values were not useful.

Examples of students scoring a *three*. Student 52 (CTSR = 14) had all the variables necessary to determine the environmental impact of cloth versus disposable diapers. Over 20 variables were used, including emissions of carbon monoxide, volatile organic compounds, nitrogen oxides, sulfur oxides, particulate matter (all from the electricity to run the washing machine and heat the water) compared across three washing machines to find the environmental cost of the cloth option, versus the emissions associated with manufacture of the cellulose, polyethylene, and adhesive of the cloth diapers.

Variable integration was scored much like variable selection. Examples of students scoring a *zero.* These students simply did not have formulas connecting variables. Student 17 (CTSR = 5), when comparing various ways to heat a house, had

data for carbon dioxide per tree, but it was not evident whether this was carbon dioxide absorbed by the tree during its lifetime or emitted from the tree when burned. While this variable seemed to be the major variable for the wood-burning side of the comparison, it was not apparently linked to the other variables. Student 27 (CTSR = 19) had data on several Energy Star appliances and regular appliances for a home, but did not do anything further with these pieces of data.

Examples of students scoring a *one.* These students had formulas, but they were wrong. Students 40 and 22 (CTSR = 17 for each) both looked at burning versus burying of trash, and both had seriously flawed formulas, such as adding the cost to build the facility to each ton of waste processed, formulas that referenced blank cells, formulas that added together cells with different units, and cells containing quantities such as cost of energy used that should have referenced the cells' dollars per unit energy and units of energy used, but did not. Student 10 (CTSR = 10) made similar mistakes, adding the cost of one single cubic yard of dirt to the yearly cost (for only one year, not a lifecycle number of years) of maintaining a landfill to a startup cost (to build a landfill) to a cleanup cost (for a whole landfill) to an operating cost per acre (which would be redundant with several other of these costs). Furthermore, none of these were calculated on a per ton basis, so that the landfill could be compared to the incinerator on a common unit (impact per ton of garbage disposed) basis.

Example of a student scoring a *two*. Student 25 (CTSR = 12) is a good example of a student scoring a *two* for variable integration. After constructing a nearly perfect model comparing the emissions from driving versus flying, student 25 then multiplied the emissions per passenger by the number of passengers, to arrive at the emissions per

vehicle, concluding that the plane emits more than the car, but granting, "However there are some positives to flying. Planes carry a lot more passengers than cars," which was what this model is supposed to factor out.

Example of a student scoring a *three*. Student 46 (CTSR = 17) connected the variables (both environmental and economic) for manufacturing, purchasing, and powering incandescent and compact fluorescent light bulbs, with no calculation errors.

Scoring how student *Checked model against data* was more difficult than scoring *variable selection* or *variable integration.* While the first two aspects of the final project rubric indicate clear relationships between CTSR and performance on a particular part of model building, the data obtained from the final projects regarding the students' abilities to check a model against data was less conclusive.

Unlike variable selection and variable integration, checking the model against data was a requirement for which success did not rest entirely with the student. For both variable selection and integration, students were urged to brainstorm variables both before and after researching the topic, to look for relationships, and in both cases, if precise data could not be found, to make reasonable estimates. Thus, success on these two tasks did not depend on finding a particular source. However, for students to compare their model against the data, they did need to have a particular type of source. Ideally, in addition to the sources that they used to gather their information, they would have been able to find a source that made a similar comparison or calculation to their model. For instance, Ask Pablo is a website that does back-of-the-envelope calculations about environmental issues, such as whether or not plastic bottles or aluminum cans are a better environmental choice for packaging beverages. Pablo is fairly forthright and clear

in his variable assumptions and calculations, and thus provides an excellent comparison for student models because they can look at both how his (albeit static) calculations were performed and also at the conclusion he reached. On the other hand, Pablo did not examine all the topics that the students used in their models, so some students had to look elsewhere.

Students looking elsewhere often found comparisons which were not as good as Pablo's comparisons. Often, the conclusion of the comparison was presented (A is better than B) with only qualitative treatment (or none at all) of the variables used to reach this conclusion. Obviously, this check of the model was not as desirable as a source that gave quantitative calculations, but it was still better than no comparison.

No comparison at all was a challenge that some students faced. Student 28 (CTSR = 19) chose to compare the environmental damage from traditional logging to that of helicopter logging. While his sources reported on variables such as fuel use and collateral damage to other trees per tree harvested, it did not reach a conclusion about which was better for the environment overall, and with such a specialized and unique topic, he was not able to find other sources with which to compare. Thus, student 28 and others were faced with little opportunity to check their model.

Another problem arose if students only found one source, such as the Ask Pablo source. If they built their model using Pablo's exact assumptions and calculations, they were only able to examine if using the same numbers and formulas resulted in reaching the same numeric answer and conclusion as Pablo did, which is far more limited than using six sources to build the model and testing the answer from the model against a seventh, independent source.

The combination of all of these complications made scoring this aspect of the project difficult. Students of all abilities, including student 53 (CTSR = 24), failed to make any explicit comparison at all between the results of their model and another's results. Rubric scores at each of the lower levels (*zero*, *one*, and *two*), indicating at least some problem with checking the model against data, encompassed the full range of CTSR scores.

Obviously, no examples of students scoring a *zero* can be given, because by definition, the student made no comparison.

Examples of a student scoring a *one.* Student 24 (CTSR = 8) said, "I compared my model to that of another student in the class who hypothesized whether frozen carrots were better than fresh, shipped carrots, and found that we had similar answers. Our numbers for both the frozen and shipped produce were pretty much the same." However, the student did not explain how or why these results were the same, based on similarities or differences in the model.

Example of a student scoring a *two*. Student 35 (CTSR = 11) said

> Comparing with another model for hybrid and non-hybrid cars, the results match but the numbers are higher for my model. One of my classmates has conducted a similar model to find out the total miles where the total pollution (environmental costs) from hybrid-electric vehicles equals the pollution from gasoline vehicles. His result is 21,550 miles which is very low compared to my model. As I mentioned earlier, this should be the result of SO2, costly gas emission factor that comes from the Nickel extraction which is included in this model … Another reason for this deviation could be the kind of cars that are compared.

As compared to the score of a *one*, above, this comparison had more details regarding why the models behaved differently.

Examples of students scoring a *three.* Student 21 (CTSR = 16) used many references to check the accuracy of the conclusions made by the model after it was completed.

> After looking at several articles, Natural versus Artificial Turf – a natural choice, from the DLF Trifolium Seeds & Science department, the NJEA article of Grass Playing Fields vs. synthetic turf, and Synthetic Turf, Health Debate Takes Root from the EHP (Environmental Health Perspectives), I have found that my conclusion is supported … all three articles they had substantial data leading to the fact the natural grass is better then (*sic*) artificial turf.

Student 21 used multiple external comparisons and claimed to have examined the data used in these comparisons. It would have been better had the student been more specific regarding the exact variables used.

The rubric for *Hypothesis testing* was again on a scale of *zero* to *three*, with a score of *zero* indicating that the student did not form a hypothesis at all. Despite its explicit mention in the directions, fully 14 of the 58 students turning in this assignment (or just under 25%) did not form a hypothesis at all. Another 15 formed a hypothesis, but it made no reference to the model at all. Thus, exactly half of the students did not use their models to form a hypothesis. By far the largest group of students (23 of 58) formed a trivial hypothesis. For the purpose of this study, a trivial hypothesis was defined as a mere extension of the original intent of the model. From the student directions:

> For example, if your model was paper versus plastic bags, how many pounds of CO_2 or units of energy would be saved by mandating a switch to using only the better bag? This type of hypothesis will be considered a trivial hypothesis because it follows directly from the model, if your output predicts that a paper bag saves $.03 over a plastic bag, then if 10,000,000,000 bags are used in the United States in a year, one only needs to multiply the above numbers to find a savings.

In fact, only 10% of students formed a hypothesis that clearly demonstrated full formal reasoning. The directions again specifically stated:

A more interesting hypothesis would be to consider how changes in your input variables would affect the output (for instance, if your model was created three years ago with gas under $2.00/gallon, does the answer change if the price of gas goes up to $3.30/gallon?) Another alternative would be to explore what value of a variable would be necessary to reverse your decision? What is the necessary price for a barrel of crude oil before plastic bags are the better option? At what price of landfill space does the option which produces the most garbage cease to be the cheapest option? What value must be assigned to a tree before the using of that tree as raw material becomes more expensive than leaving it in place to provide shade, provide CO_2 sequestration, prevent soil erosion, and other services?

Each of these paths represents another step in modeling and abstraction, to think about the input variables not in terms of what is, but in terms of what may be. Despite these instructions, only six students completed a hypothesis in which they predicted the effect of the change of at least one variable on the outcome of their model.

Student four (CTSR =5, score of *zero*) revealed a depth of misunderstanding not present elsewhere in the spreadsheet or paper when using the initial conclusion to form a new hypothesis. While the conclusion itself was fine (wood is cleaner than coal-source electricity for heating a house), student four revealed a desire to include variables that would make this comparison potentially invalid. For example, if student four multiplied the number of people using each heating source by emissions per heating source, as indicated, this would tell which source had a bigger total impact, but not which option is a better individual choice. The final conclusion here, too, is not model based, when student four begins to discuss the emissions of wood being more localized than emissions from coal-source electricity, writing,

> I think I would have tried to add the numbers of people who are actually using wood to the numbers using electricity as a heat source. I initially began looking for that number but could not find a concrete number in regards to the amount of people using wood heat. My educated guess is that the number of people using wood heat is a lot less than those who use electric. My final conclusion is that heating with electricity affects the environment on a more macrocosmic level. I believe with wood heating the pollution is felt on a localized level. In that those living inside the house with a a (*sic*) wood stove and the surrounding environment are adversely affected.

And when trying to form a hypothesis about heating a commercial building with larger square footage, student four wrote,

> That is based on a square footage of 1,000. That would mean that according to my model to heat a building of that size it would take an increase in Btu's of over

> 17 times the original amount. This would mean enormous increases in all emissions from Carbon Monoxide to Carbon Dioxide. Based on the results I would say that heating with wood in a commercial building is certainly not a viable option. Heating with electricity is more efficient in a building of that size because it can be regulated in a simple manner.

While it is not possible to know exactly what the student was thinking, it appears that the student did not realize the amount of electricity needed in the larger building would also increase, and that if wood heat released fewer emissions than coal-source electricity for a 1000 square foot house, then, all other variables being equal, a larger building should show the same advantage in the same proportions. Most other students receiving a *zero* simply failed to make a hypothesis at all.

Example of a student scoring a *one.* Student 19 (CTSR = 11, score of *one*) wrote, "A change in policy that would have a positive effect on my estimated carbon dioxide emissions would be mandating a recycle facility in every city or town to collect recyclables. This change would be a great way to reduce some of the carbon dioxide that I took into consideration on my model." Since the model had nothing to do with the distribution of facilities, and since there was no way to incorporate this policy change into the existing model, this hypothesis was not truly based on the model built, so scored a *one*.

Student 18 (CTSR = 19, score of *two*) was able to use the model constructed to think deeply about the comparison of preserving local produce versus shipping it during the offseason, saying, "From here, I went on to determine how many quarts of each method would need to be used in order for the total cost to be even. I found that when 159 quarts are

canned, 550 quarts are frozen, and 117 quarts are imported the total cost is about $100." While not as good as a level *three* hypothesis, it still uses the model to make a prediction, in this case, a breakeven between the various approaches where the larger up-front costs (environmental) of the canner or refrigerator are weighed against the larger per unit cost (environmental) of the truck.

Student 41 (CTSR = 24, score of *three*) calculated the payback time of insulation in terms of the embodied energy to make the insulation versus the energy saved over time by using the insulation.

> My hypothesis that I decided to test was to see if changes in where the energy to produce insulation came from could make any impact on the amount of carbon emissions saved. If insulation production was done using all energy from coal, the dirtiest energy source … how long would it take to make the production cost in carbon dioxide emissions to be equal to the amount of carbon dioxide emissions prevented through the installation of the insulation?

After creating the said hypothesis, student 41 tests the impact that the resultant change in energy mix (compared to the current energy mix) had on the payback time.

The *level of the model* was also assessed on a scale from *zero* to *three*, with *zero* being a non-model (a table reporting static calculations, for example), *one* being a model with only concrete components (such as the tangible objects, miles, gallons, dollars, etc.), and *two* being a model with abstract or invisible components that should be familiar (such as molecules of carbon dioxide, the environmental cost of a tree, etc.). A level *three* model was not expected, but contain postulated components or combined components in a way that is outside the typical established relationships. Lawson describes a true

scientific model such as a Mendel's gene model or Dalton's atomic model as such a model. The existence of an unknown, postulated structure with specific characteristics was necessary in each case, even though there was no direct evidence that such objects existed. While it was not surprising that no student created a new scientific model from scratch, this level would have allowed for students to, say, create a new variable if they saw the need for one. Specifically, student 37 (CTSR = 19) suggested such an approach during the human population lab. Although the HDI statistic adequately explained many trends in human development, this student suggested a better indicator might be one involving not the countries mean Gross National Product per capita, but rather the percent of the population above a threshold income. Thus, while it turned out that no student's model was deemed a level *three* in this study, it was not beyond the ability of the students in this study to postulate and create a new statistic or variable to answer a question.

Student 55 (CTSR = 4, score of *zero*) had identified several important variables comparing the energy use of boarding schools versus commuter schools. In addition, a few tentative connections between these variables were explored. However, these were reported as a static, incomplete table, not a model.

Several of these are discussed elsewhere, but these models (scoring a one) included fixation on concrete variables, such as size of a tank, miles driven, etc., as was the case with student 56 (CTSR = 13).

The bulk of students scored a t*wo*, as most models incorporated environmental costs (cost per ton of pollution), embodied energy (energy needed to make one kilogram of a substance), and/or ideas about emissions from the energy used in the comparison. For example, student 26 (CTSR = 13) calculated carbon dioxide emissions as well as the

environmental cost for both manufacturing and driving an internal-combustion gasoline vehicle, as well as a hybrid-electric vehicle.

No student scored a t*hree*, but an example of what a *three* might have consisted of has been given previously.

In addition to the five specific rubric scores already discussed, three additional variables were categorized after looking at the projects. While somewhat redundant with each of the above variables, in some ways they also simplify some of the major issues. Each will briefly be mentioned.

Although somewhat redundant with *variable selection* and *variable integration*, a separate category called *used variables* was created. This categorization was used after it was realized that a substantial percent of students did not submit a final project that was a manipulatable model and was instead a static table. Thus this category is related to variable integration in some respects, but could be different in the case that some students clearly understood how the variables could be related, but did not integrate the formulas into their spreadsheet in such a way that the changing of one number changed the output.

Another category that emerged was called *used ratios.* This categorization was used because some students did not seem to incorporate ratios into their final models. Every number used in the creating of their spreadsheet was a raw, concrete number such as tons of emissions, cost of a landfill, gallons of gasoline, etc. instead of emissions per gallon, BTU per ton, etc.

A final trend that seemed to emerge was how the students *reported final answer* at the end of the paper. Since no requirement on how the final answer was to be reported was made in the directions, this trend emerged from the students themselves.

Additionally, since some projects examined a payback time (for installing an energy-efficient product such as insulation or for building an alternative energy source such as a windmill) this categorization was not as applicable to these questions as to most others. The scoring on this category was *zero* (student said A was better than B), *one* (student said A was X units better than B), or *two* (student said A was X times or percent better than B). Obviously, students at level *two* show thinking that is more formal in nature (by examining a ratio), but also more useful. For instance, if one only knows that option A produced 10 fewer kg of carbon dioxide than option B, in order to tell if this is a meaningful difference one must know the raw amounts of carbon dioxide produced by each option as well (a 10 kg reduction would be meaningful to a process requiring 20 kg, but not to a process requiring 1000 kg). On the other hand, if one knows that option A produces half the carbon dioxide as option B, then not only can a person tell that this is a meaningful difference, but can also tell the impact to variables downstream (dollar value of carbon tax) and perhaps even upstream (amount of fuel used in the process) as these other variables are also proportionally linked to the variable in question. This simple appreciation for how the other variables will change cannot be achieved with an answer in the form of a raw difference, nor with one that only reports that one choice is better than another is.

Student 18 (CTSR = 19) best exemplified this type of thinking, stating, "However, upon thinking it through a little more, I determined that when merely comparing the different methods in terms of ratios, it doesn't matter what the costs of these pollutants are, as long as they are constant across the board," and later, "The most overall cost efficient process by a landslide is to buy locally grown strawberries, and

freeze them for six months throughout the winter. It is 5.6 times more cost efficient than transporting strawberries from California, and is 3.6 times more cost efficient than canning."

To this point no attention has been given to the two students who did not complete the project. The final modeling project, perhaps because of its weight in the class (10% of the class grade between the spreadsheet and the accompanying paper), had only two students, student 31 (CTSR = 9) and student 47 (CTSR = 10) who did not complete the project, which was the highest participation rate of any assignment except the pretest and posttest. As always, the question is whether or not these students were mentally unable to complete the project or did not complete it for other reasons (from emails, it appears that there certainly were personal issues at work in one case, but not the other).

While only two students did not complete a project at all, a number of students did not complete a spreadsheet *model* (as indicated previously when examining the individual categories). The most significant of these students to this study were the students who turned in a static chart in an Excel spreadsheet. While the variables selected by the student for the chart may or may not have been appropriate, in the end, what they submitted was not a thinking tool that allowed the user to manipulate variables and predict outcomes. The same question arises as to whether or not these students did not correctly complete the assignment because they were not able to grasp the idea of a model, or for some other reason. However, with the amount of instruction and feedback that was provided to students on what the spreadsheet was supposed to be able to do (respond to changes in input with cells that were to be linked by formulas wherever

appropriate), in many cases it seems likely that the students who submitted a static chart may have been unable to construct a model. These students had a universally low CTSR score (mean 8.17) which would be consistent with an inability to think formally. Their results are summarized in table 26.

A second group contains projects that looked like models but were completely non-functional. Members of this group had a mean CTSR score of 11.43. While these models were marginally better than the static tables submitted by the first group of students (described in the previous paragraph) because these models at least contained formulas attempting to relate the various quantities, the formulas used were either so syntactically or conceptually flawed that they did not serve their purpose. For instance, a typical mistake was made by students 22 and 10 (CTSR = 17 and 10, respectively), who both tried to combine variables that were not alike. In particular, student 10 tried to add the cost of a single cubic yard of dirt to a yearly cost (for a landfill) to a startup cost (for a landfill) to a cleanup cost (for a closed landfill) to an operating cost per acre (for a landfill). Some of these variables might overlap (the cost of a cubic yard of soil might be a component in the yearly and operating cost, and the yearly cost should be the operating cost per acre multiplied by the number of acres). The correct combination of some of these variables (startup cost plus cleanup cost plus yearly cost for the lifetime of the landfill) would be a useful number. However, adding one unit of a marginal cost to

Table 26. Students not completing an acceptable final project spreadsheet, not a model.

Student Number	*CTSR Total*	*Topic*	*What was submitted*
31	9	None	Nothing at all, personal issues given as an excuse.
47	10	Population growth	Table of exponential population growth, no manipulatable variables.
5	14	Worm composting	Table of exponential worm population growth, no manipulatable variables.
17	5	Heat: Wood vs. electricity	Static table of statistics on wood heat emissions versus electricity.
55	4	Busing vs. residential schools	No connection of variables. Did not have a common unit of comparison (per student, for instance).
14	7	Local frozen produce vs. shipped fresh produce	Only two formulas were used, and these were not correct.

several other one-time costs is not appropriate. Student 22's mistakes were more fundamental, such as subtracting the tons of carbon dioxide from the environmental cost (a completely different unit) and adding the full fixed cost of building the waste facility

to every ton of waste processed by that facility (instead of either dividing that cost amongst the total tons processed at that facility in its lifetime and adding this to the per ton cost, or conversely, multiplying the total tonnage by the per ton cost and adding that to the cost of building the facility and arriving at an accurate lifecycle cost).

Another common flaw involved what was termed (in feedback to students) the "apples to apples" comparison issue. Several students made unequal or unproductive comparisons. For instance, instead of comparing gasoline and diesel vehicles on emissions per mile driven or total emissions for the same trip, student 15 (CTSR = 11) compared the emissions of the two vehicles based on the emissions created from combusting a single tank of gas. Since the student never corrected for the differences in miles driven on a tank of gas, the model cannot yield a reliable answer. Students 29 and 19 (CTSR = 14 and 11, respectively) made similar errors. Student 29 attempted to compare the total emissions for the entire aluminum can industry to the entire glass bottle industry, without taking into account the vast difference in the number of containers of each type produced. Student 19 attempted to compare emissions from a single virgin aluminum facility to emissions from the recycling efforts of a single city. Again, this student did not consider that the number of cans produced should factor into this decision. Two final students (students seven and three, with CTSR = 9 and 8, respectively) probably were guilty of similar errors; however, these students' other errors (such as lack of appropriate units or falsified data) made positive identification of this misconception difficult. The factor that ties these errors together is that they are primarily focused on a much more concrete statistic (such as total emissions from a factory or number of gallons in a gasoline tank) rather than the more abstract (but useful)

statistic built on a ratio of two concrete statistics (such as the number of emissions *per* mile driven, or the amount of emissions released *per* ton of aluminum recycled). Table 27 summarizes this information.

There is one other group of students who did not complete an acceptable model, and this group had characteristics far different from the first two. These students typically turned in a good initial list of variables and achieved a good start to the model (linking some of these variables appropriately), but did not complete a finished model. Student 49's (CTSR = 19) failure to complete a model seemed to stem from selecting too many variables (over 75!) to research and integrate, and the student ran out of time.

Student 37 (CTSR = 19) had difficulty finding a specific data point, and stopped when that data was not found, but appeared to be headed in the correct direction. No clarification was provided by student in the third case, student 40 (CTSR = 17). On average, these three students had good CTSR scores (mean 18.33). The students who submitted an incomplete project are detailed in Table 28.

Pretest and posttest analysis

The pretest and posttest were intended to be the primary quantitative measure of student gain in understanding of models and nature across the course. However, as explained elsewhere, their emerged a split with regard to modeling, with the ability to actually construct a model being more fully assessed by the final modeling project, and with the pretest and posttest serving as summative assessment for the understanding of scientific models and the nature of science.

Table 27. Students not completing an acceptable final project spreadsheet, fatally flawed model.

Student Number	*CTSR Total*	*Topic*	*What was submitted*
15	11	Diesel vs. gasoline	Comparison of emissions per tank full of gas rather than a meaningful comparison of emissions per mile.
22	17	Landfill vs. incineration	Mixed up real and environmental costs, added costs to tons of emissions directly, added the cost of the facility to each and every ton of waste processed.
29	14	Glass bottles vs. aluminum cans	Wanted to compare pollution totals on an industry vs. industry basis (ignoring the vast difference in units produced) instead of pollution per unit.
19	11	Aluminum recycling vs. virgin	Tried to compare recycling in one city vs. virgin production for a company, on a total emissions basis rather than a per container basis.
7	9	Aluminum vs. glass vs. plastic containers	Errors of multiplying when the student should have divided, confused capacity (oz.) of the object with weight (oz.) of the object, bizarre units like "emissions for glass" in "millions of ounces per km." Not clear if is emission was for the industry as a whole or one bottle?

Table 27. Continued.

Student Number	*CTSR Total*	*Topic*	*What was submitted*
10	10	Incinerator vs. landfill	Did not compare on a per ton basis. Added the cost of 1 cubic yard of dirt to a yearly cost to a startup cost to a cleanup cost (for a whole landfill) to an operating cost per acre, so what was done made no sense.
3	8	Tire derived fuel vs. recycling tires.	Not "apples to apples". Unable to deal with the ratios of energy saved and emissions saved by one process vs. energy spent and emissions created by another. Did not arrive at a final answer. Could not grasp emission per unit energy ratio. Made up data?

Table 28. Students not completing an acceptable final project spreadsheet, submitted incomplete spreadsheet.

Student Number	*CTSR Total*	*Topic*	*What was submitted as the final project*
20	17	Landfill vs. incineration vs. recycling	Model was not complete. Correct variables were identified, but actual data used was "made up" (student admission) and formulas were more than conceptually wrong, they referenced empty cells, etc…
49	19	Nuclear vs. coal electricity	Model not complete. What was complete was more than most students turned in, and was correct in data and relationship, but this student's attempt to include every aspect made the comparison too complex to reasonably complete in the time allotted.
37	19	Improving energy efficiency of appliances vs. house itself.	Model not complete. Was a table of calculations of payback times for energy efficiency upgrades (appliances, insulation, alternative energy source such as wind/solar). Email explanation indicated difficulty in finding some data.

The instrument, a modified version of the SUMS combined with a modified version of the SUSSI, collected student responses to Likert scale as well as free response questions regarding the models and the nature of science. Each type of question presented its own strengths and weaknesses. In particular, a handful of students left the posttest free response questions blank, resulting in loss from pretest to posttest. However, the free response answers gave potentially much better insight into what the students actually thought. On the other hand, the Likert-scale questions were never unanswered, but there were some issues with students trying to outsmart the test that were possible with Likert-scale questions that were not possible with free response.

A sub-score for various aspects of the nature of science and modeling was created, using a sum of all questions pertaining to that particular aspect. For nature of science questions, the SUSSI provided the structure for which questions to combine and for modeling, the SUMS provided that structure. In a few cases with the SUMS, interaction with students in the follow-up interviews indicated that students perceived these questions in a way other than the intent. Because of this, a few questions have been included in more than one sub-score.

Overall, the sub-score categories are listed in Table 29.

Question 39 does not appear in the above table as it represents a standalone misconception relating to multiple models and the educational construct of learning styles. The categories *uses/purposes of scientific models* and *multiple models* had

Table 29. Sub-score categories and component question

Category	*Questions*
Nature of Science	1-6
Theory Change	7-11
Multiple models	13, 26, 27, 28, 29, 30, 31, 32
Explanatory tools	16, 17, 18, 21, 28
Exact replicas	16, 19, 33, 34, 35, 36, 37, 38
Uses/purposes of scientific models	13, 14, 20, 21, 22, 28, 29
Changing nature of models	23, 24, 25
Types of models	12
How are models created	15, 36
Scientific method(s)	40-44

significant overlap of questions (13, 21, 28, and 29) because one of the reasons that multiple models of the same phenomenon exist is to fulfill different purposes such as explaining different aspects of the phenomenon in question. For instance, the Lewis Dot Structure of an atom is a model that can explain how an atom bonds, but tells nothing either about the nucleus of the atom or the three dimensional shape of a molecule. On the other hand, a Bohr model of the atom gives more detail about inner electrons shells and the nucleus, but is cumbersome to use in bonding compared to the Lewis Dot Structure. Valence Shell Electron Pair Repulsion models show three-dimensional representations of molecules, but do not show multiple bonds like a Lewis Dot Structure can. Thus, the very fact that atoms have many interesting behaviors requires multiple models to answer

specific different questions about their behavior. Likewise, since the *use/purpose of* many *scientific models* is as an *explanatory tool*, there are some questions (21 and 28) that are also deemed to fall in both of these categories as well.

Another instance where there is an overlap is between e*xplanatory tools* and e*xact replicas*, as these questions are two sides of the same coin. If a student answers disagree to a question like Question 16, (Likert Scale) which said ”Scientific models are only used to physically or visually represent something,” it should be because they think that a scientific model is not an exact replica and is primarily used as an explanatory tool.

Question 36 (Likert-scale), which asked students to agree or disagree with the statement “All parts of a model should have an understandable purpose/reason,” seemed to apply as equally to how a model was created as to the original SUMS classification for models as exact replicas.

Each of the sub-scores in question will be examined in more detail in the following section.

Analysis of gains on each question and sub-score.

For each of the questions on the pretest/posttest, an analysis of the change in the scores will follow. For Likert-scale questions, analysis will be more limited, looking at any other trends not present above in the correlations and a tentative explanation as to why this change was observed. For each free response question, a more detailed analysis detailing statistics (such as word counts) of pretest and posttest answers and how these statistics changed will be analyzed, again, with tentative explanations for any trends observed.

Question one. Likert-scale. *Nature of hypotheses, theories, and laws.* Value, one point. Question text: *"*Scientific theories exist in the natural world and are uncovered through scientific investigations." Correct answer: S.D. Rationale: Theories are created by scientists to explain the natural world. Normalized change (-0.20). The mode normalized change was zero, (with 24 students not changing their answers), with the most common normalized change being -1, from 18 students who answered *agree* on the pretest and *strongly agree* on the posttest, resulting in a normalized change of -1. The 13 students showing a positive change were outnumbered almost 2:1 by the 23 students showing a negative change. This question provided difficulties for students as reflected in the follow up interviews with students. For each of the students in the interview, the word *uncovered* appeared to be the word that caused the difficulty because they did not understand what exactly was meant. For students participating in the interview, the question was reworded as follows. "There is a classic riddle regarding the whether or not a tree makes a sound when it falls in the woods, if no one is there to hear it. Is a theory like that sound, existing in nature and waiting for a human to observe it, or is the scientist's role in the theory more active than mere observation?" Students agreed that this wording seemed to explain what *uncovered* meant, and in some cases caused students to change their answers. It was expected that students would show gains on this question, because of student's experience making models during the course of the semester, and the similarity of model creation to theory creation, but this was not observed.

Question two. Likert-scale. *Nature of hypotheses, theories, and laws.* Value, one point. Question text: "Unlike theories, scientific laws are not subject to change." Correct answer: S.D. Rationale: Scientific laws are subject to change, for instance Newton's

Laws of motion do not hold at relativistic speeds (although students have less experience with laws changing than theories, as most examples are found in modern physics). Follow-up interviews revealed no particular misunderstandings with the wording of this question; the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (25 students) and the most common was again a normalized change of -1 (10 students), from a change of answer from agree to strongly agree. Students changing to a more incorrect response outnumbered students changing to a more correct response 20 to 15. Since the only laws discussed in class were the laws of conservation of matter and conservation of energy and no explicit discussion of this process occurred, it is not surprising that this question did not show a large change from pretest to posttest.

Question three. Likert-scale. *Nature of hypotheses, theories, and laws.* Value, one point. Question text: "Unlike theories, scientific laws are not subject to change." Correct answer: S.D. Rationale: Theories and laws answer different questions. Laws tell what phenomenon will be observed, often with great accuracy, but theories postulate why. Follow-up interviews revealed no particular misunderstandings with the wording of this question; the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (30 students) and the most common change was again a normalized change of -1 (19 students), from a change of answer from agree to strongly agree. Students changing to a more incorrect response outnumbered students changing to a more correct response 20 to 10. Since the only laws discussed in class were the laws of conservation of matter and conservation of energy and the only theory discussed was global warming, the class did not lend itself to explicitly

teaching these relationships and it is not a concept that would follow from the activities completed.

Question four. Likert-scale. *Nature of hypotheses, theories, and laws.* Value, one point. Question text: "Scientific theories explain scientific laws." Correct answer: S.A. Rationale: The Kinetic-Molecular Theory explains the various gas laws. The Theory of Relativity explains Newton's Law of Universal Gravitation. Follow-up interviews revealed no particular misunderstandings with the wording of this question; the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (21 students) and the most common change was a normalized change of -0.66 (12 students), from a change of answer from agree to disagree. Students changing to a more correct response outnumbered students changing to a more incorrect response 21 to 18. Since the only laws discussed in class were the laws of conservation of matter and conservation of energy and the only theory discussed was global warming, the class did not lend itself to explicitly teaching these relationships and this concept is also not a concept that would follow from the activities completed.

Question five. Likert-scale. *Nature of hypotheses, theories, and laws.* Value, one point. Question text: "Scientific theories are hypotheses that have been tested many times and not disproven." Correct answer: S.A. Rationale: Some hypotheses become theories through repeated testing. Follow-up interviews revealed no particular misunderstandings with the wording of this question, the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (30 students) and the most common change was a normalized change of one (nine students), from a change of answer from *agree* to the correct answer of *strongly agree*.

Students changing to a more correct response outnumbered students changing to a more incorrect response 16 to 14. There was some discussion during the lecture on global warming about how an observation of a trend can lead to a hypothesis and then to a theory, however, this was only a small part of one lecture. Although students created hypotheses, they did not get to see hypotheses become theories or laws. The class did not lend itself to explicitly teaching these relationships and this concept is also not a concept that would follow from the activities completed.

Question six. Free-response. *Nature of hypotheses, theories, and laws.* Value, three points. Question text: "With examples where appropriate, what is the nature (definition) of each: law, hypothesis, and theory. Then, explicitly state the differences and relationships between each." Correct answer: Hypotheses are predictions, based on scientific knowledge, about the outcome of an experiment. A law is a scientific statement, often mathematical, generally regarded as true. A theory is an overarching explanation of a set of related observations or events. Hypotheses that are supported can become parts of theories or laws. Theories do not become laws, contrary to student beliefs, but may explain them. There were no misunderstandings of the wording of this question; students appeared to be answering the question incorrectly due to legitimate misunderstandings.

Several trends were evident throughout the pretest and posttest answers to question six, with more similarities than differences between pretest and posttest answers from the same student when examined side by side. In other words, word counts and other analysis support the idea that student conceptions of the definitions of and

relationships between laws, theories, and hypotheses appeared to have changed less than expected over the course of the 15 week class.

This question also posed perhaps the most difficulty in scoring, as it attempted to ascertain too much: three definitions and up to three pairwise relationships. Thus two identical scores could represent completely different scenarios, such as good definitions but poor relationships, good relationships but poor definitions, or any combination in between. These scenarios necessitated a closer, side by side examination of the actual responses to ascertain which areas improved or failed to improve.

Vocabulary also proved to be a difficulty. As the data in Table 30 shows, certain words such as *true, proven, unable to be changed, 100% correct* are used repeatedly by students. In general, all of these statements would be considered to be too strong when applied to the concepts of theories and laws. Both theories and laws are very well supported, and have not been proven wrong yet, but a scientists would stop short of saying a law has been proven. Is this student use of *proven* when they mean *supported* semantics, or do students truly believe that laws cannot change? To further complicate this issue, students occasionally made statements such as "Proven though testing" or "proven many times with the same results" or "a hypothesis has been proven many times without fault". Although these statements are still incorrect because of the word *proven*, they do show that the student understands that hypotheses that are tested and turn out to be correct lend support to laws and theories, and that consistent results are necessary to move a hypothesis towards becoming a theory or a law. However, student responses to the corresponding and less ambiguous Likert-scale questions pertaining to theories,

Table 30. Word Count in answers to question six on the pretest and posttest

Word	*Number of times it appears in the pretest*	*Number of times it appears in the posttest*	*Notes*
Prove, proven, etc.	108 in 3399 words	81 in 3419 words	25% decrease in frequency,
True	35	30	
Correct	10	12	
Total Prove+True+Correct	153	123	20% decrease in frequency.
Explain	12	28	133% increase in frequency
Model	0	4	Used correctly.

hypotheses and laws supports the idea that, in general, students tend to think that laws have been *proven 100% true* and *correct* and are *unable to be changed*..

Students were not even self-consistent within a single paragraph answer, for instance, implying laws can never be changed ("doesn't waiver") but also that they are hard to change, and thus could change ("don't usually [change] … so easily"). Another student gave a similarly ambiguous answer regarding theories. "…Theory - A hypothesis that has been tested and proved true … a theory can be disproven …" Overall, answers that were inconsistent were not given the benefit of the doubt and were scored based on the incorrect rather than the correct portion.

In a side by side analysis of answers, 45% of student answers on the posttest question six showed at least some improvement over answers to the same question on the

pretest, 46.3% of student answers either showed no change or were inconclusive, and 8.7% of answers on the posttest were worse than the pretest answer from the same student, as shown in Table 31. *Vastly improved* (or *vastly worse*) was a change of more than one full point. *Improved* (or *worse*) was a change of one point. *Slightly improved* (or *slightly worse*), was a change significant enough to notice, but not enough to result in a different rubric score. There are several fundamental student misconceptions that appeared to be resistant to instruction which could explain the lack of gain observed.

As a whole, improvement seemed to be related to a better understanding of the word theory and its relationship to law and hypothesis. As was predicted, since models and theories are closely related and the method of this study was using model modification to understand theory modification, it would be expected that most improvement would occur with models. However, the amount of improvement overall was quite small (only 19 students of 60 or 31.6% improved or vastly improved their answer to question six), indicating that students still held onto many misconceptions.

The most prevalent misconceptions surround the word *theory.* There are many possible misconceptions relating to the word *theory*. First there is the lay definition of the word *theory* that is different from that of the scientific definition. In everyday language *theory* is synonymous with hunch or idea, in science, theory is one of the (relatively few) large, well supported, overarching ideas that organizes scientific thought within a discipline. The importance and level of support involved with the scientific definition is obviously much greater than the lay definition, which does not require any basis of support. In fact, a lay theory does not even need the level of background or logic

Table 31. Comparison of quality of posttest answer to pretest answer on question six

Result	Number	Percent
Vastly improved	7	11.7
Improved	12	20.0
Slightly improved	8	13.3
Total Improved	**27**	**45.0**
No change	23	38.3
Inconclusive	5	8.3
Total Inconclusive + NC	**28**	**46.7**
Slightly worse	4	6.7
Worse	1	1.7
Total worse	**5**	**8.3**

necessary to form a good scientific hypothesis (i.e. an educated guess). A second issue with the word theory is that some theories including the Theory of Evolution and the Big Bang Theory are highly controversial in some segments of the population. It appears possible that students and teachers may deal with this controversy by demoting theories to a lower level of importance. Both of these explanations are consistent with the data that was observed in question six, and summarized in Table 32.

Specific data that supports the above assertions include:

- Denying theory's power to explain phenomena and ascribing it instead to laws (a law, such as the Ideal Gas Law PV=nRT, may accurately predict how a gas will behave, but offers no explanation to why the gas behaves

in this manner). This occurred four times each on the pretest and posttest.

- Relegating theories to a level below hypotheses, for instance "Hypothesis: An Educated Guess that you Think Is Correct, Theory: A Guess you Hope is correct, A Theory Is A Hypothesis you don't Know Anything about." This misconception appeared five times on the pretest but did not appear on the posttest, indicating an area where students improved their understanding of the nature of science.

- Using the lay definition of theory. For example "A theory is a plausible explaination [*sic*]." The frequency of this misconception decreased from four instances on the pretest to two instances on the posttest.

- "Hypothesis is an idea for example evolution." Or "Sometimes, in the case of evolution, it will remain a hypothesis simply because we cannot truly know what took place." Here, the most powerful idea in biology, the Theory of Evolution, is reduces to a mere "idea".

- "A law is something that has been proven and therefor *[sic]* is true. Examples are the laws of gravity." Or "Law: Something that is. EX: Law of Gravity." Or "Laws are widely accepted as in gravity." Or finally "A law is a a [*sic]* theory that is held as true and is not disproven, for example the law of gravity."

Table 32. Concept Count in answers to question six on the pretest and posttest

Concept	*Number of times it appears in the pretest*	*Number of times it appears in the posttest*	*Notes*
Hierarchy	34	32	4 "implied" cases in each.
Law and theory reversed or the (incorrect) idea that laws explain	4	4	Typically, granting the power of explanation to laws
Theory and hypothesis reversed	5	0	Typically, stating a theory is just a guess and can become a hypothesis through testing.

Table 32. Continued.

Confusion about order of hypothesis and data collection. Hypothesis not related to experimentation. Confusing hypothesis with purpose or observation.	2	4
Theory is just a guess/lay definition of theory	4	2
Lay definition of law	1	1

Related to the various misconceptions regarding the word *theory* is the relationships between *theory, hypothesis* and *law*. The correct scientific understanding of the relationship of these three words is that some hypotheses, when tested and supported with data may become laws. Theories too, come from hypotheses which have accurately predicted outcomes. Laws tend to come from hypotheses about what will occur, theories come from hypotheses about underlying reasons why something occurs. Furthermore, a theory is used to generate further hypotheses, which if successful, will further support the theory. Thus, a schematic for this relationship might look like Figure 5.

However, 34 students on the pretest and 32 students on the posttest (more than half in both cases) described a different relationship between hypotheses, theories and laws, as shown in Figure 6.

This much more linear, hierarchical relationship relegates theories to a position between laws and hypotheses. While this relationship does accurately depict the frequency with which each of these ideas may be changed or revised (hypotheses the most, laws the least, although laws are occasionally revised such as relativistic versions of Newton's Laws of Motion) it is otherwise inaccurate most specifically because it depicts theories becoming laws.

Furthermore, there are additional implied misconceptions in this hierarchical structure. One misconception is that a theory that is well supported becomes a law, thus only theories that are not well supported stays at the theory stage. This interpretation is consistent with explanations above attempting to denigrate theories, especially controversial ones, to a lower status. Finally, such a schematic implies a certain temporal misconception that is counter to historical facts relating to laws and theories concerning the same phenomenon. According to the hierarchical schematic, a theory must predate a corresponding law. However, there are many examples where this order is obviously not the case. Boyle's Law, which predicts the behavior of the volume of a gas under pressure, predates the Kinetic Molecular Theory (which adequately explains why pressure increases as volume decreases) by many decades (1662 vs. 1734). Moreover, the Kinetic Molecular Theory encompasses not only the Boyles Law, but also all of the other gas laws and concepts such as diffusion. Newton's Law of Universal Gravitation predates the Theory of General Relativity that explains gravity by centuries. However, despite these obvious examples, students still hold on to this hierarchical view, even after instruction relating to global warming theory as a way of explaining previously observed trends, observations, etc.

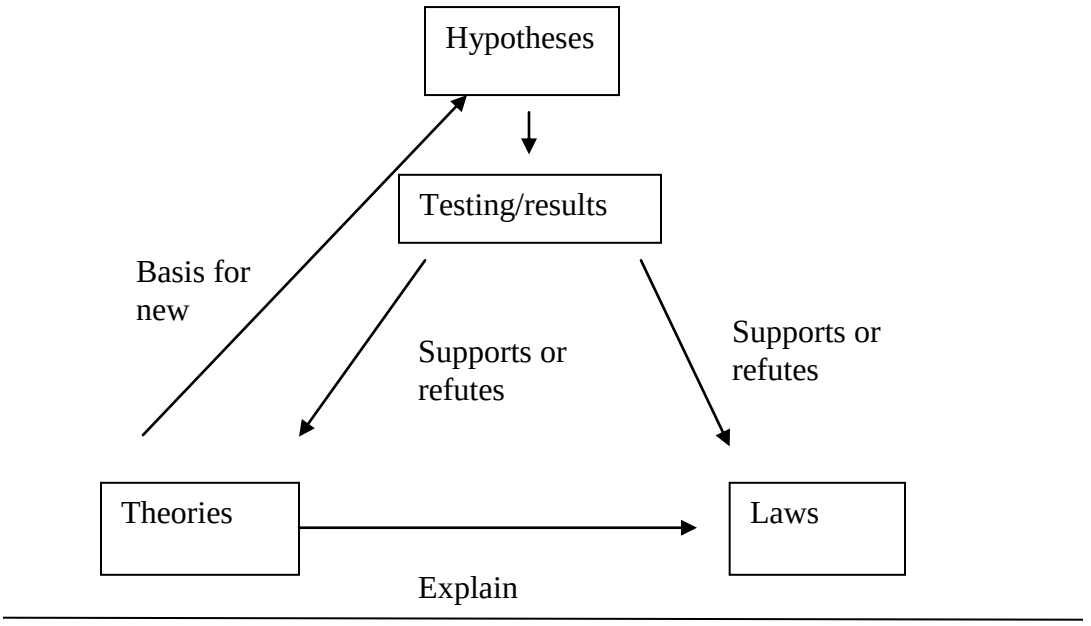


Figure 5. Correct, parallel conception of key science concepts.

As discussed previously, students received low scores on the pretest and/or posttest because they tended to describe laws as being proven. This idea is contrary to the scientific conception. As shown in Table 30, the number of times words synonymous with *proven* were used was 153 on the pretest, and 123 on the posttest. While this decrease represents approximately a 20% decrease, there are still a substantial number of students who see laws as being proven facts.

On the other hand, there was a marked increase (133%) in the number of students using the word *explain* on the posttest as compared to the pretest. This supports the idea that student conceptions of theories improved, as theories are responsible for explaining phenomena.

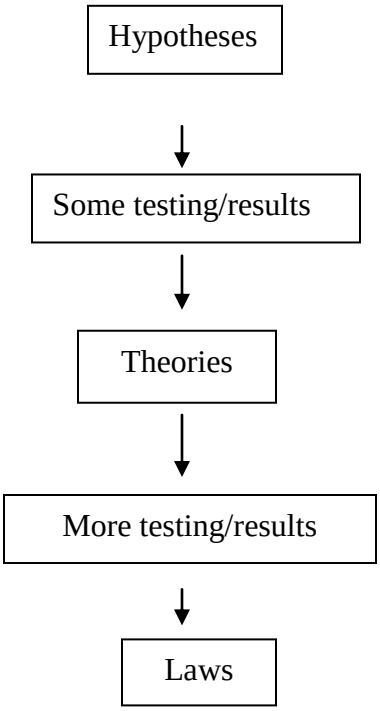


Figure 6. Hierarchical (and incorrect) conception of key science concepts

On the posttest, the word *model* appeared four times in answers to question six. It was used correctly in its relation to both hypothesis and theory. For instance, "Theory: … attempts to explain the model," and "Theory: a testable model." This relationship between theory and model underlies the methodology of the study, unfortunately, only two students specifically referenced this relationship in their answers to question six.

While it was expected that students would show gain in their understanding of the nature of science, specifically, the meaning of laws, theories, and hypotheses and the relationship between these concepts, gains were small

Summary of questions one through six. Questions one through five tended to show a negative normalized change not because of any meaningful change from *agree* to

disagree but because of shifts in degree (from *agree* to *strongly agree*) which may have represented real change in the strength of conviction or may have been gaming the system (several students spoke of using this strategy of answering only strongly agree or strongly disagree on the posttest). This lack of gain is reflected in the effect sizes (Cohen's *d*) of 0.09, 0.12, 0.21., 0.22 and 0.31. Question six, although much more difficult to score, did provide very concrete evidence of gains, with almost one third (19/60) improving their answer by a full point and almost half (27/60) showing some improvement, and only five giving a worse answer. These gains were primarily due to a decrease in regarding theories and laws as completely correct and eternal, and in increase in the notion of theories for explanation

Question seven. Likert-scale. *Theory change* sub-score. Value, one point. Question text: "Scientific theories are subject to on-going testing and revision." Correct answer: S.A. Rationale: Theories are modified over time, for instance, many of the parts of Dalton's Atomic Theory are no longer correct due to discoveries of the subatomic particles and isotopes. Follow-up interviews revealed no particular misunderstandings with the wording of this question, the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (42 students) and the most common change was a normalized change of one (15 students), from a change to the correct answer of *strongly agree*. Students changing to a more correct response outnumbered students changing to a more incorrect response 16 to two. Because of the extensive revision of models that occurred during this class, it was hypothesized that students would improve their understanding of how theories change.

However, since nearly three-quarters of students did not change their answer, the improvement seen (0.23 normalized change) was not large.

Question eight. Likert-scale. *Theory change* sub-score. Value, one point. Question text: "Scientific theories may be completely replaced by new theories in light of new evidence." Correct answer: S.A. Rationale: Theories are modified over time, for instance, the heliocentric theory of the solar system completely replaced the geocentric theory, as the heliocentric theory provided a better explanation for the observed behavior of the planets. Follow-up interviews revealed no particular misunderstandings with the wording of this question, the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (42 students) and the most common change was a normalized change of one (14 students), from a change to the correct answer of *strongly agree*. Students changing to a more correct response outnumbered students changing to a more incorrect response 16 to two. Because of the extensive revision of models that occurred during this class, it was hypothesized that students would improve their understanding of how theories change. However, since nearly three-quarters of students did not change their answer, the improvement seen (0.23 normalized change) is again not large.

Question nine. Likert-scale. *Theory change* sub-score. Value, one point. Question text: "Scientific theories may be changed because scientists reinterpret existing observations." Correct answer: S.A. Follow-up interviews revealed no particular misunderstandings with the wording of this question, the misunderstandings revealed by the test seem legitimate. The most common result from pretest to posttest was no change (38 students) and the most common change was a normalized change of one (11

students), from a change to the correct answer of *strongly agree*. Students changing to a more correct response outnumbered students changing to a more incorrect response 18 to four. Because of the extensive revision of models that occurred during this class, it was hypothesized that students would improve their understanding of how theories change. However, since nearly three-quarters of students did not change their answer, the improvement seen (0.21 normalized change) is again not large.

Question ten. Likert-scale. *Theory change* sub-score. Value, one point. Question text: "Scientific theories based on accurate experimentation will not be changed." Correct answer: S.D. Rationale: A theory may correctly explain all *accurate* experimentation that exist at that time, and yet still be changed as new data becomes available. Follow-up interviews revealed no particular misunderstandings with the wording of this question, the misunderstandings revealed by the test seem legitimate, although it should be noted that students had difficulty with the concepts of this question. To them, the whole process of theory development and change was somewhat vague, and without a good understanding of the process, it is difficult to envision scenarios under which a theory built on accurate experiments would change, particularly as this specific example was not discussed in class. The most common result from pretest to posttest was no change (35 students) and the most common change was a normalized change of 0.66 (seven students), from a change from the incorrect answer *agree* to the more correct answer of *disagree*. Students changing to a more correct response outnumbered students changing to a more incorrect response by the narrow margin of 14 to 11. Since more than twice as many students did not change their answer as improved it, and almost as many

students' answer became more incorrect on the posttest, the improvement was minimal (normalized change 0.04).

Question eleven. Free-response. *Theory change* sub-score. Value, three points. Question text: "Do scientific theories change? If yes – how (in what ways and to what extent) and why? If no – why not?" Correct answer: Yes. The theories may change gradually or radically based on new evidence.

Overall, some improvement in student's understanding of the changing nature of theories was achieved. At worst, 23 (and perhaps as many at 25) of the 60 students showed at least some improvement in this answer from pretest to posttest, with most showing improvement of a full point due to conveying a better understanding that theories may change is both small and large ways. In addition, 28 to 30 showed no change from pretest to posttest, and only seven students had worse answers on the posttest than the pretest, with the bulk of these students failing to indicate the extent to which a theory can be changed on the posttest after having explicitly done so on the pretest. This data is summarized and presented in Table 33.

One inconclusive score was at least *no change* and perhaps a gain of some sort. One aspect of the understanding models was explicitly articulated in the pretest, but missing from the posttest but a different aspect was explicitly articulated in the posttest, but not the pretest. This results in "no change". However, the use of the word "information" in the posttest but not the pretest was unclear. Was this "information" a synonym for data or evidence (in which case this demonstrates an improvement)? The other inconclusive score showed an improvement in the posttest over the pretest in

recognizing that evidence was needed to change a theory, however, it also specifically mentioned the misconception that such evidence would change a theory into a law. This misconception is severe enough it was decided not to give credit for this improvement in recognizing the relationship between new evidence and changing of theories.

Summary of question seven through 11. Overall, an effect size of 0.36 was achieved on question 11. This result seems consistent with the results on the corresponding Likert scale questions, which showed effect size on average of 0.15, with two questions with much larger effect sizes (Cohen's d = 0.3 and 0.31). These questions related to reinterpreting existing data and complete revision of theories, respectively. As Table 33 shows, these gains clearly match the specific reasons for increases in student scores from pretest to posttest.

Question 12. Free-response. *Types of models* sub-score. Value, three points. Question text: "List as many scientific models as you can." Correct answer: A variety of models should be represented including physical, mathematical, and conceptual/theoretical models.

Question 12 was another straightforward question to score. This question showed large gain, with 38 of 60 students having a more complete answer on the posttest than on the pretest, and only 21 students with unchanged answers and only one student scoring lower on the posttest than the pretest. A closer examination of the students who improved their scores reveals that 25 of the 38 improved by adding one category, 11 students added two categories, and two students had a posttest answer with the all three categories of models after having no models listed correctly in the pretest.

Table 33. Concept count in answers to question eleven on the pretest and posttest

Pretest to posttest result	*Number*	*Percent*	*Reason*
Vastly improved (2 points)	1	1.7	Clarified extent of change and the reason for change (data/evidence)
Improved (1 point)	16	26.7	2 added "evidence" (or synonym) in posttest answer 12 added extent of change in posttest answer 1 add different way of thinking about existing data to posttest 1 switched from theories may not to theories may change
Slightly Improved	6	10.0	3 Theories may change because of new ways of thinking (posttest only) in addition to new evidence (present in both). 1* Some indication of extent of change present in posttest but not full extent of changes possible. 2* Became more specific, using "evidence" in the posttest instead of "information" in the pretest.

Table 33. Continued.

Pretest to posttest result	*Number*	*Percent*	*Reason*
Slightly Improved, continued	6	10.0	1 Eliminated wrong answer from pretest to posttest, specifically that a theory cannot completely change.
No change	28	46.7	Pretest and posttest were essentially the same.
Slightly worse	2	3.3	1 Explicit wrong statement about theories in posttest (but not related to rubric). 1 Answer is less explicit about extent of change in posttest, but provides examples that show extent.
Worse	5	8.3	4 had indicated extent in pretest but not in posttest. 1 had theories may change in pretest, but may not in posttest.
Inconclusive	2	3.3	See below.

* Star indicates a student who *slightly improved* in 2 areas.

Of the 25 students whose scores improved by one level, 13 specifically improved by adding mathematical examples from class to their posttest answer. Furthermore, 36 of the 60 posttest answers made direct reference to activities in class, and math models were mentioned 44 more times on the posttest compared to the pretest. Thus, much of this gain appears to be a direct result of the activities of the class.

The students had multiple, multiday experiences with mathematical models (Resource Lab, Carbon Footprint Lab, Human Population Lab, and individual final Excel projects) but relatively less experience with conceptual (only two, a simulation that visualizes global warming and food webs/chains in lecture) and no experiences with physical models in this class. Student answers shifted away from physical models (a decrease from 25 answers mentioning physical models on the pretest to only 23 answers mentioning physical models on the posttest) towards mathematical models (an increase from 10 answers mentioning mathematical models in the pretest to 54 (representing 90% of all responses) mentioning mathematical models in the posttest. However, the fact that 12 of these 25 who improved one level, plus the 13 students who improved by more than one level were able to add models to their posttest that were not the specific mathematical models used in class shows that perhaps some general knowledge of models had been transferred. In addition, conceptual models increased to a smaller extent, from 17 answers mentioning conceptual models on the pretest to 26 answers mentioning conceptual models on the posttest, so not all gain came from mentioning activities directly from class.

Question 12 represents the entire types of model sub-score itself, as *types of models* was not a part of the original SUMS question

Question 13. Free-response. *Multiple Models* sub-score. Value, three points. Question text: "Multiple models exist of the same phenomenon, such as a map of the United States. Why?" Correct answer: Different models reflect different aspects of the phenomenon in question (roads, political boundaries, geography) of the same phenomenon (the United States). Each model serves a different purpose. Follow-up

interviews revealed no misunderstandings with wording. Furthermore, this question was relatively easy to score, with few questions that required interpretation on the part of the scorer. These questions were flagged as *vague*, and the benefit of the doubt was not given with respect to scoring.

Overall, this question showed the strongest gains, as not a single student's response on the posttest was worse than their response on the pretest, with 16 students showing no change and 44 students showing some measureable gains, with 28 improving by one point and 11 improving by two or more points.

More detailed analyses of specific concept frequency and word frequency data are presented in Tables 34 and 35 respectively. There were many specific concepts and words indicative of an understanding of multiple models which appeared with greater frequency in the posttest than the pretest. Regarding concepts, these included the complexity/accuracy tradeoff, the purposes/uses of models, and the central concept of this question, that a phenomenon has many different aspects and a model captures only some of these. In addition, a decrease was seen in the number of answers labeled vague, and a total disappearance of the misconception that multiple models of the same phenomenon cannot exist because then at least one would have to be wrong.

Table 35 revealed similar information regarding word count. While some of the words used seemed to indicate a level one or level two conceptualization of a model (to see what something looks like or to show/teach/communicate about the model) increasing dramatically, there was also an increase in the number of students who discussed the level three concepts of interpretation. Overall, then, it can be said that posttest answers

Table 34. Concept counts on question 13.

Concept	Pretest count	Posttest count	Comment
Complexity/ accuracy	0	2 complexity, 3 accuracy, 2 with both	This reflected an appreciation for the idea that a real phenomenon is often too complex, and may require simplification.
Purpose, uses	3,4	11,4	This increase shows an appreciation for the purposive nature of models. Models are created with a specific purpose/uses in mind.
Aspects	19	38	Different models of the same phenomenon typically reflect different aspects of that phenomenon. This increase represents most of the gain.
Vague	9	4	Student answers became less vague and more precise.
One model is wrong	4	0	Students often hold a misconception that more than one model can exist only if one model is wrong. This misconception, although not widely displayed in the pretest, was not present in the posttest.

Table 35. Word counts on question 13.

Word	*Pretest count*	*Posttest count*	*Comment*
Aspect	7	17	Aspects implies that different behaviors of a phenomenon, as opposed to different physical sides.
Show	10	46	*Show* can mean to communicate to another, a higher level of understanding than merely visually see a replica.
See, look	3, 7	10, 28	This increase is troubling as *see* and *look* are words associated with visual models. However, *look* was often used in another way, such as to *look at one aspect or another*.
Interpret	6	8	This increase is consistent with models for understanding.

were more complete than the pretest answers and captured all of the ways models can be used.

Question 14. Free-response. *Use/purpose of models* sub-score. Value, three points. Question text: "What is the most important characteristic of a scientific model or, in other words, what characteristic makes a scientific model the most useful? Explain." Correct answer: The ability to make accurate, testable hypotheses and to adequately explain a variety of observations. One difficulty in analyzing this question lies in the student's use of the word *variable*, particularly in posttest answers. A second difficulty

stems from the word *accurate* and its derivatives. This question was difficult to score because student use of some very key words was unclear, and follow-up interviews only further reinforced the idea that different students could mean to very different ideas by using the same words.

The word *variable* poses a problem in scoring because it may or may not show a high level of understanding. The word *variable* does not exist in the rubric, because a variable is literally a function of only a mathematical model, and the question is concerning models in general. However, taken more broadly, a variable is that which is manipulated. In a physical model of an atom with Velcro electrons to stick on an off, the electrons are that which is manipulated, and thus in some very general sense might be considered a *variable*. In any type of model, however, the usefulness of the model centers around conclusions that can be made when that which is manipulated is in fact manipulated. Moreover, the correct use of variables is one of the primary indications of formal thought. It would be very helpful at this point to further query student who answered *variables* to see what they would say the most important aspect of a non-mathematical model was, however, data collection has ended.

The use of the word *variable* in an answer does not indicate that a particular student understands *variables*. Students of all cognitive abilities used the word *variable* in their answer; there was no correlation with cognitive ability. Classroom examples, however, did not show all students were equal in their understanding of the usefulness of variables in a model. Drafts of models and interactions with students during model building revealed a number of students who initially created an Excel spreadsheet that did not correctly use variables, as described in detail in the previous section.

Considering two extreme examples, some instances where the word *variable* is used seem to indicate more of a level one understanding of models, reflecting an increased level of detail, or a more exact replica. On the other hand, most examples seem to allude to a level two understanding of models stressing ideas such as functional similarities, or even a level three conception emphasizing accurate predictions, without explicitly stating it.

The word *accuracy* poses similar problems, especially when presented without other context. Does the word imply *accurate predictions* (level three understanding of models on the rubric) or more merely more detail, as an exact replica (level one understanding of models on the rubric). Initially, this dilemma was a source of inter-rater reliability error as the other scorer focused more on the word than the overall meaning in context. Unfortunately, in two cases there was no context, just the single word "accuracy," and several other cases where the context is very limited and did not clarify what the student meant. These have uniformly been scored at a level one. There is a final problem regarding accuracy. Several answers have attached the idea of accuracy specifically to the input values/data. While it is important to strive for accurate inputs and accurate relationships in a model, a model is only an approximation of reality. As an approximation, it can never be completely accurate (a common model misconception, in fact) and it is completely acceptable in models to use variables and relationships that are good enough to provide accurate predictions, even if more accurate representations are available, particularly if using such less accurate inputs reduces the complexity of the model. This idea of accurate inputs is not explicitly reflected in the rubric, except possibly in terms of detail. These answers as well will be scored at a level

one, as it would appear they probably are related more to the misconceptions regarding a model as an exact replica.

Even though scoring was difficult and a number of scores which potentially could have been much higher were reduced to a score of one, results regarding the hypotheses for question 14 showed some gain. The average score on the pretest was 1.13, and the average score on the posttest was 1.81 yielding a large effect size using *Cohen's d* of 0.72. Table 36 captures many of the ways that these scores improved.

The most common reason a student was classified as vastly improved was an indication of understanding that scientists use models to form predictions or hypotheses. The most common reasons a student was classified as improved were moving from no answer or a completely incorrect answer to a physical model understanding (level zero to level one, which happened in six instances) and moving from an idea of a model for explaining to a model for predicting (level two to level three, six instances).

Of the students who scored worse on the posttest, a majority of them gave an answer that was more specific regarding the activities in class than their original pretest answer, which was more generally applicable (three students). It was also interesting to note that three students scored a perfect three on the pretest, only to score lower on the posttest.

Analysis of the word counts for question 14 supported the trends observed elsewhere. These specific word counts can be seen in Table 37. As with the previous question, there was a decrease in language (such as *prove*) which might indicate a feeling for what *visually* meant, a mathematical formula can decrease in misconceptions

Table 36. Changes in students' answers on question 14 from pretest to posttest

Difference between posttest and pretest	*Number of students showing that difference*
Posttest answer was vastly improved from the pretest (an increase of 2 levels or more):	15 students
Posttest answer was improved from the pretest (typically an increase of 1 level):	15 students
Posttest answer was slightly improved from the pretest (typically an more complete answer at the same level):	6 students
No change between pretest and posttest answer:	18 students
Posttest answer was worse than the pretest:	8 students
Inconclusive (typically, one or more of the responses was ambiguous enough to make drawing a conclusion difficult) :	10 students

regarding models being exact replicas. There was also a large increase in the use of the words related to *variables*, likely because few students conceived of mathematical models before the class and most of the models in class were mathematical in nature. There was also an increase of language related to making predictions and hypotheses. One student specifically mentioned quantification as a purpose of models.

Question 15. Free-response. *How are models made* sub-score. Value, three points. Question text: "A headline reads 'Global warming model predicts sea-level will

Table 37. Selected word counts for question 14.

Word	*Pre*	*Post*
Total visual/physical language	8	12
"Vari" as in variable	0	27
"Predict" or "hypothesis"	4	10
"Prove"	7	4

rise 2 meters by 2100 A.D.' What do they mean by "model" and how was this model created?" Correct answer: This mathematical model was likely physically created on a computer by conscious choice of the variables and data to include and omit. Students seemed to have a difficult time understanding this question. The primary source of confusion was rather than focusing on the cognitive aspect of how the model was created (variable selection and integration) students focused on the technical aspects (with a computer) and if they felt uncertain about these technical aspects, tended not to answer the question, particularly on the pretest.

While the pretest version of question 15 was a good question (A headline reads "Global warming model predicts that sea level will rise 2 meters by 2100 AD". What do they mean by "model" and how was this model created?) it was not used on the posttest in the exact same form, because of the fact that this phenomenon (a global warming model) was used in great depth in class. Use of the same question would perhaps allow students to show gain because of memorizing a specific answer of how the global warming model viewed in class was made, rather than understanding how models are constructed in general. A different version of this question appeared on the Summer and

Fall posttests version (A headline reads "EPA models show that raising the average miles per gallon of U.S. vehicles by 1 mpg would reduce gasoline supply/demand pressure better than drilling in the Arctic National Wildlife Refuge." What do they likely mean by "model" and how was this model created?). This question was slightly different in that students were less likely to be able to score a one easily by saying “trend” or score a two by saying “extend a trend” and this question really forces students to talk about variable selection and relation to earn points. As a result of this slightly more difficult and focused nature of the posttest question, the gains seen should be a result of a better understanding of what is needed to make a functioning model in general and as likely come as a result of memorization of ideas presented in lecture. Overall, question 15 showed gains with *Cohen’s d = 0.89* and a normalized change of 0.40, reflective of the specific gains mentioned in Table 38.

Question 16. Likert-scale. *Exact replicas* and *explanatory tools* sub-scores. Value, one point. Question text: “Scientific models are only used to physically or visually represent something.” Correct answer: S.D. Rationale: Mathematical, conceptual, or theoretical models may be non-physical or non-visual. Follow-up interviews revealed some severe issues with this particular question. The words *only*, *or* and *visually* were the source of confusion. Some models are visual or physical, which made students think that the answer was true, however, not all models are. In addition, students did not have a strong feeling for what visually meant, a mathematical formula can be seen, therefore, is it not visual? was their argument. Having one problematic word in this question was bad enough, but with all three it is not surprising that the results do not show any improvement.

Table 38. Analysis of pretest/posttest trends in question 15.

Result	*Number*	*Comment*
Vastly improved (improved by two)	10	Selection of specific variables, manipulation of variables are commonly added to the posttest.
Improved (improved by one point)	26	Movement from a vague "something to do with a computer" to an understanding of specific process.
Slightly improved (not enough to change score)	6	Minor clarification, slightly more explicit, answer otherwise similar
Total improved	42	70% of students improved their score from pretest to posttest for question 15.
No change (12) + inconclusive	13	21.7% of students did not change their score from pretest to posttest

Table 38. Continued.

Result	*Number*	*Comment*
Total students having a worse (three students, one point worse) or much worse (two students, two points worse)	5	8.3% of students received a worse score on the posttest than on the pretest for question 15. Three posttest answers were too vague or tangential to the question asked that they could not be scored. One mentioned specific variables, but they were incorrect variables. Four answers clearly related to concepts covered in class, which made them too specific.

The most common result from pretest to posttest was no change (21 students) and the most common change was a normalized change of one (11 students), from a change to the correct answer of *strongly disagree*. Next most frequent (nine students) was a change from agree to disagree, which shows some students made the progress towards the acceptable answer. Students changing to a more correct response outnumbered students changing to a more incorrect response 25 to 14. Because of the extensive use of non-physical models during this class, it was hypothesized that students would improve their understanding of non-physical nature of models. The improvement seen (0.12

normalized change, 0.07 effect size) is not supportive on gains in knowledge regarding physical and visual models. Thus, while originally categorize in the *explanatory tools* category, and re-categorized by the author to go in the *exact replica* category, perhaps this question would have been best off deleted entirely.

Question 17. Likert-scale. *Explanatory tools* sub-score. Value, one point. Question text: "Scientific models are used to explain scientific phenomena." Correct answer: S.A. Rationale: The JAVA Climate Change model that students worked on in class attempted to explain the relationship between fuel use, population growth, greenhouse gas concentrations and temperature. The Lewis Dot Structure of an atom explains ionic and covalent bonding under simple conditions. Follow-up interviews revealed only a slight apparent issue with the wording of this question, and that was *phenomena*. The definition of this word was provided on the pretest and posttest to any student who asked, however.

The most common result from pretest to posttest was no change (33 students) and the most common change was a normalized change of one (18 students), from a change to the correct answer of *strongly agree*. Next most frequent (four students) was a change from disagree to agree, which shows some students made the progress towards the acceptable answer by completely reversing their views. Students changing to a more correct response outnumbered students changing to a more incorrect response 22 to five. Because students were asked to make hypotheses and to build explanations from several models throughout the semester, they should have been comfortable with the idea that a model could explain. The improvement seen (0.29 normalized change, 0.17 effect size) support some growth in student understanding of the idea of models as *explanatory tools.*

Question 18. Likert-scale. *Explanatory tools* sub-score. Value, one point. Question text: "Scientific models may be used to show an idea." Correct answer: S.A. Rationale: Same as question 17. Follow-up interviews revealed no apparent issues with the wording of this question, although a few students had a harder time with the concept of *showing an idea* than explaining a phenomenon, once they understood what a phenomenon was.

The most common result from pretest to posttest was no change (33 students) and the most common change was a normalized change of one (22 students), from a change to the correct answer of *strongly agree*. These two outcomes, between them, represented all but five of the students. Students changing to a more correct response outnumbered students changing to a more incorrect response 25 to two. Because students were asked to make hypotheses and to build explanations from several models throughout the semester, they should have been comfortable with the idea that a model could explain some rather abstract concepts, such as a carbon footprint. The improvement seen (0.38 normalized change, 0.45 effect size) support some solid growth in student understanding of the idea of models as *explanatory tools.*

Question 19. Likert-scale. *Exact replicas* sub-score. Value, one point. Question text: "A scientific model is a diagram, picture, map, graph or photo of a physical object." Correct answer: S.D. Rationale: Most models used in science for investigations are not of physical objects, but rather of relationships. This question posed a problem in the pilot study, and was revised. Follow-up interviews revealed that there were still difficulties, primarily due to the words *or* and *physical object*. Several of these items (certainly maps are models, graphs are mathematical models showing a relationship, and diagrams such

as a food web or Kreb's cycle are scientific models) can be models. A photo, however, is not considered a model. Furthermore, many scientific models are of concepts not physical objects. However, there was enough confusion with this question that it could have been removed from the analysis.

The most common result from pretest to posttest was no change (21 students) and the most common change was a normalized change of -1 (18 students), from a change to the incorrect answer of *strongly agree*. On the other hand, eight students reversed (correctly) from *agree* to *disagree*, with only three students reversing the other way. Students changing to a more incorrect response outnumbered students changing to a more correct response narrowly, 22 to 17. The lack of improvement seen (-0.16 normalized change,- 0.04 effect size) may support confusion with the question, as opposed to a lack of understanding of models.

Question 20. Likert-scale. *Uses/purposes of models* sub-score. Value, one point. Question text: "Models are used to help formulate ideas and theories about scientific events." Correct answer: S.A. Rationale: As has been discussed elsewhere, models and theories are synonymous in science. There appeared to be no difficulty understanding the question during the follow-up interviews.

The most common result from pretest to posttest was a change to the correct answer (29 students) followed closely by no change (28 students), with only three students showing a different outcome and only two students making a more incorrect answer on the posttest than the pretest. Students changing to a more correct response outnumbered students changing to a more incorrect response 30 to two. As would be

expected based on these results, the statistics yielded good gain (0.47 normalized change, 0.74 effect size) on the central idea of this question. This result is not surprising, as students were asked to formulate hypotheses using the models on each of the modeling activities.

Question 21. Likert-scale. *Uses/purposes of models* and *models as explanatory tools* sub-scores. Value, one point. Question text: "Scientific models' primary value is in showing/teaching science." Correct answer: S.D. Rationale: Models' primary value lies in their ability to make accurate predictions on behavior, at their heart, scientific models are thinking tools. Follow-up interviews spent a fair amount of time on this question, particularly when students answered questions 20 and 22 correctly but question 21 incorrectly. The question was posed to students in the interview what the professors in the science building used models for, most answered teaching students. It was then pointed out that these scientists do research when not teaching class, and do these professors use models when acting as a scientist. Some remarked it was not necessary, since scientists already know the material, and most of these students displayed an overall ignorance of what a scientist does. Others made comments relating to communication with their colleagues. Very few thought scientists might use models to make hypotheses. Therefore, the results of this question are probably slightly influenced by this ignorance of what a scientist does, and if students better understood scientific endeavor, they might have answered more correctly on this question.

This confusion aside, the most common result from pretest to posttest was no change (25 students) followed a change to the completely wrong answer (10 students) and a change from agree to the more correct disagree (10 students). Students changing to

a more correct response outnumbered students changing to a more incorrect response 20 to 15. As would be expected based on these results, the statistics yielded virtually no gain (0.01 normalized change, 0.19 effect size).

Question 22. Likert-scale. *Uses/purposes of models* sub-score. Value, one point. Question text: "Models are used to make and test predictions about a scientific event." Correct answer: S.A. Rationale: Models' primary value lies in their ability to make accurate predictions on behavior, at their heart, scientific models are thinking tools. Follow-up interviews revealed no misunderstanding about this question.

The most common result from pretest to posttest was no change (30 students) followed a change to the completely correct answer (25 students). Students changing to a more correct response outnumbered students changing to a more incorrect response 27 to three. As would be expected based on these results, the statistics showed large gains (0.41 normalized change, 0.63 effect size). Taken together, questions 20 – 22 show that although students seem very comfortable with the use of models as thinking tools, they could not get over their misconceptions that the primary purpose of scientific models is teaching.

Question 23. Likert-scale. *Changing nature of models* sub-scores. Value, one point. Question text: "A model can change if new theories or evidence prove otherwise." Correct answer: S.A. Rationale: Models, like theories, change with new evidence. Follow-up interviews revealed no confusion on this question.

The most common result from pretest to posttest was no change (46 students) followed a change to the completely correct answer (13 students), with the one remaining

student moving from neutral to agree. This question suffered somewhat from ceiling effect, as the average score on this question was 0.85 on the pretest. Statistical analysis showed moderate gains (0.22 normalized change, 0.41 effect size).

Question 24. Likert-scale. *Changing nature of models* sub-score. Value, one point. Question text: "Once created, a model does not change." Correct answer: S.D. Rationale: Models, like theories, change with new evidence. Follow-up interviews revealed no confusion on this question. Like its sister question 23, this question too suffered somewhat from ceiling effect, with a 0.85 on the pretest.

The most common result from pretest to posttest was no change (42 students) with all other students changing to the completely correct answer (18 students). Statistical analysis showed moderate gains (0.30 normalized change, 0.37 effect size).

Question 25. Likert-scale. *Changing nature of models* sub-score. Value, one point. Question text: "A model can change if there are changes in data or beliefs." Correct answer: S.A. Rationale: Models, like theories, change with new evidence. Follow-up interviews revealed no confusion on this question.

The most common result from pretest to posttest was no change (37 students) followed a change to the completely correct answer (19 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 21 to one. Statistical analysis showed large gains (0.34 normalized change, 0.69 effect size). Taken together, questions 23 through 25 ask essentially the same question in opposite ways, providing a test to the reliability of the results. As results were similar even though the answer had changed (strongly disagree vs. strongly agree), questions 23 through 25

give more faith in the reliability of this instrument than in the original SUMS, where each question was phrased to the positive.

Question 26. Likert-scale. *Multiple models* sub-score. Value, one point. Question text: "Multiple models of the same phenomenon/object are typically used to express features of a phenomenon/object by showing different perspectives to view/see a phenomenon/object." Correct answer: S.D. Rationale: Multiple models of the same phenomenon tend to show different interactions a phenomenon may make, rather than different views/perspectives of how an object looks. For instance, as described previously, the Lewis Dot Structure of an atom shows bonding, whereas the Bohr model of an atom shows the nucleus. Follow-up interviews on the pilot study revealed that this question was troublesome. After rewording, follow-up interviews during this study revealed that this question was still too cumbersome to be easily understood, or that students did not comprehend the ultimate importance of models not being about viewing/seeing the phenomenon/object but rather predicting its behavior.

The most common result from pretest to posttest was a change to the completely incorrect answer (30 students) followed a change no change (24 students). Students changing to a more incorrect answer outnumbered students changing to a more correct answer 32 to four. Statistical analysis showed large negative changes (-0.48 normalized change, -0.77 effect size). As almost all other questions relating to multiple models showed large gains, it would seem logical that the wording was confusing to students.

Question 27. Likert-scale. *Multiple models* sub-score. Value, one point. Question text: "Multiple models of the same phenomenon/object represent different

versions/aspects/facets of the phenomenon/object." Correct answer: S.A. Rationale: Multiple models of the same phenomenon tend to show different interactions a phenomenon may make, rather than different views/perspectives of how an object looks. This question was almost exactly like question 26, however, it differed in not having the emphasis be on merely viewing or seeing the phenomenon object. Overall, students mostly answered these two questions exactly the same, despite the difference.

The most common result from pretest to posttest was no change (29 students) followed a change to the most correct answer (26 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 27 to four. Statistical analysis showed moderate gains (0.39 normalized change, 0.46 effect size).

Question 28. Likert-scale. *Multiple models, uses/purposes of models* and models as *explanatory tools* sub-scores. Value, one point. Question text: "Models can show the relationship of ideas clearly." Correct answer: S.A. Rationale: As with earlier questions, the purpose of models is as a thinking tool, explaining relationships and behaviors. Although the SUMS initially classified this question as a multiple model question, it would appear to have little to do with multiple models. There was no apparent difficulty understanding this question as revealed by the follow-up interview.

The most common results from pretest to posttest were no change (25 students) and change to the most correct answer (25 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 33 to two. Statistical analysis showed very large gains on this question (0.48 normalized change, 1.01 effect size).

Question 29. Likert-scale. *Multiple models* and *uses/purposes of models* sub-scores. Value, one point. Question text: "Multiple models of the same phenomenon/object are used to show differences in individual's theories on what things look like and/or how they work." Correct answer: S.A. Rationale: As with earlier questions, models are synonymous with theories. Potentially, the fact that this question contained an *or* linking a visual use of models is somewhat concerning, although it did not appear to cause a problem in the follow-up interviews.

The most common result from pretest to posttest was no change (33 students) and the most common change was a change to the most correct answer (16 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 20 to seven. Statistical analysis showed small gains on this question (0.22 normalized change, 0.19 effect size).

Question 30. Likert-scale. *Multiple models* and *uses/purposes of models* sub-scores. Value, one point. Question text: "Multiple scientific models are used primarily to show different sides or shapes of an object." Correct answer: S.D. Rationale: As with earlier questions, many scientific models are not physical. Like question 26 another question trying to probe at the students' attachment to the idea of physical models, this question regarding multiple models did not show as much gain as questions 27-29. Unlike question 26, however, question 30 was much more concisely and clearly worded.

The results for question 30 were very dispersed, with answers running almost the full gamut. The most common result from pretest to posttest was no change (22) and the most common change was a change to the most correct answer (10 students). Students

changing to a more correct answer outnumbered students changing to a more incorrect answer by the narrow margin of 20 to 18. Statistical analysis showed small gains on this question (0.03 normalized change, 0.16 effect size). While small, these positive gains stand out in contrast to the very large negative changes observed in question 26, and ostensibly similar question. Therefore, it does stand to reason that a fairly large percent of the negative change shown in that question is related to the wording, rather than the concept.

Question 31. Likert-scale. *Multiple models* sub-score. Value, one point. Question text: "Multiple models of the same object/phenomenon may use different information." Correct answer: S.A. Rationale: Models showing different aspects of a phenomenon may only use information pertaining to that aspect and omit other information not pertinent to that aspect in an attempt to make the model simpler. This question probably could have been classified as a *uses/purposes of models* question as well since multiple models are typically designed with different purposes in mind, and that purpose shapes the information chosen. There appeared to be no difficulty with understanding this question during the follow–up interviews.

The most common result from pretest to posttest was no change (33 students) and the most common change was a change to the most correct answer (23 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 26 to one. Statistical analysis showed large gains on this question (0.41 normalized change, 0.77 effect size). Since students specifically examined multiple models and compared and contrasted the inputs used (particularly in the Carbon Footprint Activity) this result is consistent with expectations.

Question 32. Likert-scale. *Multiple models* sub-score. Value, one point. Question text: "A model has what is needed to show or explain a scientific phenomenon." Correct answer: S.A. Rationale: Models have what is needed, the necessary information. Sometimes, they do not have much more, for instance, the Lewis Dot Structure manages to explain many bonding interactions while ignoring the nucleus and in the case of larger atoms, the majority of the electrons in the atom. As with question 31, this question probably could have been classified as a *uses/purposes of models* question as well since multiple models are typically designed with different purposes in mind, and that purpose shapes the information chosen. There appeared to be little difficulty with understanding this question during the follow-up interviews, although *has what is needed* was a little vague.

The most common result from pretest to posttest was no change (27 students) and the most common changes were a change to the most correct answer (14 students) and from disagree (incorrect) to agree (14 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 28 to five. Statistical analysis showed large gains on this question (0.33 normalized change, 0.75 effect size). Since students specifically examined multiple models and compared and contrasted the inputs included and omitted (particularly in the Carbon Footprint Activity) strong student gains on this question were expected.

Question 33. Likert-scale. Models as *exact replicas* sub-score. Value, one point. Question text: "A scientific model should be an exact replica of the object." Correct answer: S.D. Rationale: If a model were an exact replica, it would no longer be a model, it would be the original. Of the exact replica question, this question is the most clear.

There appeared to be no difficulty with understanding this question during the follow-up interviews.

The most common result from pretest to posttest was no change (30 students) and the most common changes were a change to the most correct answer (nine students) and from agree (incorrect) to disagree (nine students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 20 to 10. Statistical analysis showed small gains on this question (0.13 normalized change, 0.17 effect size).

Question 34. Likert-scale. Models as *exact replicas* sub-score. Value, one point. Question text: "A model needs to accurately represent the object/phenomenon in the areas of interest." Correct answer: S.A. Rationale: As with question 32, this question assesses students' understanding of the idea that a model often captures a simplified representation of the phenomenon. There appeared to be no difficulty with understanding this question during the follow-up interviews.

The most common result from pretest to posttest was no change (32 students) and the most common change was a change to the most correct answer (22 students) with the remaining four students also moving towards the most correct answer. Statistical analysis showed large gains on this question (0.42 normalized change, 0.74 effect size). These results are similar to the results of question 32, which is a similar question.

Question 35. Likert-scale. Models as *exact replicas* sub-score. Value, one point. Question text: "A model should closely resemble the object/phenomenon, so nobody can disprove it." Correct answer: S.D. Rationale: The utility and thus longevity of a model depends more on its ability to functionally represent the phenomenon, not the apparent

physical similarity. Follow-up interviews revealed some issues with student understanding of this question. As has been clearly shown in previous questions (32 and 34), students understand that models need to have important aspects of the phenomenon in order to accurate. In addition, if a model or theory provides accurate predictions and explanations, it will be accepted and if it does not, it will be rejected. However, the idea that a model or theory can be made infallible by closely (physically?) resembling the object/phenomenon being modeled is where statement's truth falls apart. However, the truths are blunt, numerous, and obvious, the error is small and subtle.

The most common result from pretest to posttest was no change (31 students) and the most common change was a change from agree (incorrect) to disagree (a more correct answer) (8 students). Students changing their answer to a more incorrect answer slightly outnumbered students changing their answer to a more correct answer 15 to 14. Statistical analysis reflected these trends, showing slight negative changes (-0.06 normalized change, -0.18 effect size). These results are markedly different from the results of questions 32 and 34.

Question 36. Likert-scale. Models as *exact replicas* and *how models are created* sub-scores. Value, one point. Question text: "All parts of a model should have an understandable purpose/reason." Correct answer: S.A. Rationale: A good model has face validity. Particularly with mathematical models, the question of which variables to include and which to omit and how much to weight each variable is essential to model building. Follow-up interviews revealed no issues with student understanding of this question. This question, too, is related to previous questions (32 and 34), where students

have shown that they understand that models need to have important aspects of the phenomenon in order to accurate.

The most common result from pretest to posttest was no change (39 students) and the most common change was a change to the most correct answer (19 students), with the remaining two students also changing to a more correct answer. Statistical analysis reflected these trends, showing a large gain (0.34 normalized change, 0.71 effect size). These results are in line with the results from questions 32 and 34. This large gain is also in agreement with the large gains shown on question 15, which was the only other question relating to how models are made.

Question 37. Likert-scale. Models as *exact replicas* sub-score. Value, one point. Question text: "A scientific model needs to be close to the real thing by being very exact in every way except for size." Correct answer: S.D. Rationale: Since many scientific models are NOT physical, most are not scale models. This question, like question 35, gets at the misconception of models as exact replicas. Again, while many familiar models in science class are scale models (atoms, cells, etc.) most scientific models are not. Follow-up interviews did not reveal any concerns with the wording of the problem.

The most common result from pretest to posttest was no change (25 students) and the most common change was a change from agree (incorrect) to disagree (a more correct answer) (10 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 21 to 14. Statistical analysis reflected these trends, showing a very small gain (0.07 normalized change, 0.13 effect size).

Question 38. Likert-scale. Models as *exact replicas* sub-score. Value, one point. Question text: "A model shows what the real thing does and/or what it looks like." Correct answer: S.A. Rationale: A scientific model typically reflects the behavior of the phenomenon or object. Here, the focus on what the target looks like only because of the word *or*. Follow-up interviews did not reveal any concerns with the wording of the question.

The most common result from pretest to posttest was no change (28 students) and the most common change was a change to the most correct answer (15 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 24 to eight. Statistical analysis reflected these trends, showing small gains (0.27 normalized change, 0.33 effect size).

Question 39. Likert-scale. Not incorporated into sub-score. Value, one point. Question text: "Multiple models are important for different student learning styles." Correct answer: S.D. Rationale: Contrary to some students' beliefs, multiple models do NOT have anything to do with learning styles. A valence shell electron pair repulsion model a molecule is more visual/spatial than a quantum mechanical model of an atom, but only one can be used to determine magnetism, regardless of whether the user is visual/spatial or kinesthetic. Follow-up interviews did not reveal any concerns with the wording of the question, although students only have a vague impression of what learning styles are (and no idea whether or not that construct itself has any validity). Students clung to this misconception, although students who had some familiarity with a variety of models could be forced through a carefully structured set of examples in the follow-up interviews into realizing why their reasoning was erroneous.

The most common result from pretest to posttest was no change (31 students) and the most common change was a change to the most incorrect answer (21 students). Students changing to a more incorrect answer outnumbered students changing to a more correct answer 23 to six. Statistical analysis reflected these trends, showing a large negative change (-0.31 normalized change, -0.33 effect size). Thus, despite the fact that students answered a question during the Carbon Footprint Activity regarding how they might use the multiple models presented differently (and learning style was not mentioned), when directly asked they still felt multiple models were related to learning styles.

Question 40. Likert-scale. *Scientific method* sub-score. Value, one point. Question text: "Scientists use different types of methods to conduct scientific investigations." Correct answer: S.A. Rationale: A cancer drug researcher will adhere much more closely to the textbook scientific method, with experimental and control groups, than scientists in a more purely observational field such as astronomy or field biology. Follow-up interviews did not reveal any concerns with the wording of the question.

The most common result from pretest to posttest was no change (39 students) and the most common change was a change to the most correct answer (13 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 16 to five. Statistical analysis reflected these trends, showing small gains (0.18 normalized change, 0.03 effect size).

Question 41. Likert-scale. *Scientific method* sub-score. Value, one point. Question text: "Scientists follow the same step-by-step scientific method." Correct answer: S.D. Rationale: This question is the opposite of question 40, and thus should have the opposite answer. Question 41, unlike question 40, explicitly mentions the scientific method, which may result in the somewhat different results. Follow-up interviews did not reveal any concerns with the wording of the question.

The most common result from pretest to posttest was no change (25 students) and the most common changes were to both to the most correct answer (7 students) and to the most incorrect answer (7 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 20 to 15. Statistical analysis reflected these trends, showing small gains (0.05 normalized change, 0.20 effect size). While these results seem quite similar to the results of question 40, there is an important difference noted if gains are set aside and raw scores are examined. On the posttest for question 40, student responses averaged 0.8 out of 1.0, indicating almost every student agreed or strongly agreed that scientists in different fields use different methods. However, when the phrase *scientific method* was added, performance plummeted with an average score of only 0.51 out of 1.0. Thus, even though gains were roughly the same, it would appear that when the word *scientific method* is included, students are much more likely to believe that all scientists follow it, than if those specific words are not used.

Question 42. Likert-scale. *Scientific method* sub-score. Value, one point. Question text: "Correct use of the scientific method guarantees accurate results." Correct answer: S.D. Rationale: The scientific method does not automatically eliminate random or systematic error. Follow-up interviews did not reveal any concerns with the wording

of the question, however, it did reveal that non-science majors were not equipped to understand the nuances of this question.

The most common result from pretest to posttest was no change (31 students) and the most common changes were to both to the most correct answer (9 students) and to the most incorrect answer (9 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 18 to 11. Statistical analysis reflected these trends, showing small gains (0.06 normalized change, 0.01 effect size).

Question 43. Likert-scale. *Scientific method* sub-score. Value, one point. Question text: "Experiments are not the only means used in the development of scientific knowledge." Correct answer: S.A. Rationale: As stated in question 40, scientists in different fields have different approaches. One of those approaches is to build a model, and unlike question 40-42, this idea was explicitly addressed in class. Follow-up interviews did not reveal any concerns with the wording of the question.

The most common result from pretest to posttest was no change (39 students) and the most common change was to the most correct answer (15 students). Students changing to a more correct answer outnumbered students changing to a more incorrect answer 20 to one. Statistical analysis reflected these trends, showing small gains (0.29 normalized change, 0.08 effect size).

Question 44. Free-response. *Scientific method* sub-score. Value, three points. Question text: "With examples, explain whether scientists follow a single, universal scientific method OR use different types of methods." Correct answer: Astronomy or field biology may be purely observational of natural phenomena. Other sciences may use

strictly controlled experiments. Different methods are valid for different disciplines. Follow-up interviews did not reveal any concerns with the wording of the question; however, it did again reveal that most non-science majors (and even some science majors) had little knowledge of how science was conducted in any field, let alone between two fields.

The most common result from pretest to posttest was no change (17 students), but students changing to a more correct answer outnumbered students changing to a more incorrect answer 30 to 13. Statistical analysis reflected these trends, showing small gains (0.08 normalized change, 0.45 effect size).

While this question was a free-response question and worthy of more analysis, it was also a surprise free response question at end of a long string of Likert-scale questions. Students tended not to answer completely, and did not use examples. Furthermore, other than the idea that models are one means of scientific investigation, this question was not discussed explicitly in class and was tangentially related at best. Therefore, not further analysis was performed.

As a whole, questions 40-44 represent a section of the test that could have been omitted. Although small gain was seen in some questions, overall, this section was not closely enough related to the classroom activities to merit inclusion.

The previous pages have focused almost exclusively on gain, which is an acceptable way to measure learning, particularly if students start with a variety of initial abilities. However, sometimes it is good to know whether the gains moved students from

low failing to high failing (but still failing) or from low passing to high passing, or perhaps most importantly, from high failing to low passing.

Particularly in the case of questions regarding specific misconceptions, students did not make this jump from having misconceptions prior to the class to not having these misconceptions after class. While there is great danger in doing statistics higher than frequency counts with Likert-scale items, it is easier to read a mean of 0.8 and realize most of the students were on one side of the scale and a mean score of 0.5 shows that the students were more balanced than to look at a frequency table. Student misconceptions were still common in the posttest in questions *one, two, *three, four and six (regarding the nature of laws, hypotheses and theories), 16 (science models are physical/visual), 19 (science models of a physical object), 21 (a science model's primary value is in teaching science), *26 (multiple models are to visually see different views of the same object), 30 (science models as physical models), 35 (correct models cannot be changed), 37 (science models are scale models), *39 (multiple models and learning styles), 41 (scientific method is universal), 42 (scientific method guarantees accurate results), and 44 (scientific method is universal). Each of these questions had a class averaged lower than 60%. Those with stars were lower than 21% on the posttest, which effectively means that had any of these questions shown gains, these gains were not meaningful in the real world, as 80% of these students still held the incorrect conception.

More germane to this study, it shows that the key student misconceptions regarding the nature of science and modeling are very resistant to change.